INFORMATION-THEORETIC RESULTS ON COMMUNICATION PROBLEMS WITH FEED-FORWARD AND FEEDBACK

by

Ramji Venkataramanan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Sandeep P. Sadanandarao, Chair
Professor David L. Neuhoff
Professor Demosthenis Teneketzis
Assistant Professor Erhan Bayraktar

To my parents and grandparents.

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible.

I am indebted to my advisor Prof. Sandeep Pradhan for his expert guidance, support and encouragement. Even at times when I was frustrated with research, meetings with him always left me with new ideas and renewed optimism. His high standards in research, creativity and insistence on 'high-level' understanding of a problem are qualities I hope to emulate in my own career.

I am grateful to Prof. David Neuhoff and Prof. Demos Teneketzis for their interest in my research and for sharing their deep knowledge. I have learned a lot from both of them and am delighted to have them on my committee. It was a pleasure and privilege to have been a TA for Prof. Teneketzis. He has been a great mentor and helped me mature as a graduate student. I would also like to thank my cognate committee member, Prof. Erhan Bayraktar for his keen interest in my research problem and for sharing a perspective from a different field.

I wish to thank Prof. Anastasopoulous for two excellent courses in coding theory, and Profs. David Barrett and Michael Woodroofe for teaching me analysis and measure-theoretic probability. I learned a great deal about good teaching from the excellent instructors at Michigan. I wish to thank Becky Turanski, Nancy Goings, Ann Pace and Beth Lawson for efficiently and cheerfully helping me deal with myriad administrative matters.

I am grateful to Dr. Jonathan Yedidia of Mitsubishi Electric Research Labs for

offering me an internship. It was exciting to work in the area of iterative coding and I learned a lot from him that summer.

I have thoroughly enjoyed living in Ann Arbor and that is largely due to the wonderful friends I have had. I feel fortunate to have been part of the MATC- it gave me something to look forward to each week and I thank all my friends there. In particular, Kelly O'Doherty has been a dear friend who shared my highs and lows and been a constant source of support. I have been lucky to have had wonderful roommates- my heartfelt thanks to Vijay Srivatsan, Sandeep Tata, Manoj Rajagopalan. Ali Nazari and Dinesh Krithivasan have been great office-mates. I thank Somesh Srivastava, Aditya Mahajan and Arvind Rao for interesting conversations, especially during lunch-time. I also wish to thank my best friends from high school- we are now scattered around the world, but make sure we meet regularly.

I thank my uncle and aunt, Mr. and Mrs. Sivaraman for their warm hospitality and wonderful meals. They made my initial transition to the States so much easier. Finally, and most importantly, my heartfelt thanks to my parents, grandparents and sister, for their love and support from thousands of miles away which kept me going through the hard times.

TABLE OF CONTENTS

DEDICATION . . . . . **ii**

ACKNOWLEDGEMENTS . . . . . **iii**

LIST OF FIGURES . . . . . **vii**

LIST OF TABLES . . . . . **viii**

LIST OF APPENDICES . . . . . **ix**

ABSTRACT . . . . . **x**

CHAPTER

1. Introduction 1

1.1 Source coding with feed-forward 2
1.1.1 Contributions 5
1.2 Channel coding with feedback 6
1.2.1 Contributions 9
1.3 Evaluating the expressions 9
1.3.1 Contributions 11
1.4 Multiple Descriptions with feed-forward 11
1.4.1 Contributions 13

2. Source Coding with Feed-Forward 14

2.1 Introduction 14
2.2 The source coding model 18
2.2.1 Problem statement 18
2.2.2 Intuition behind the proposed approach 20
2.3 Stationary and ergodic joint processes 22
2.4 General sources 32
2.4.1 Rate-distortion theorem 32
2.4.2 Discrete memoryless sources 40
2.4.3 Gaussian sources with feed-forward 41
2.5 Error exponents 44
2.5.1 'Good' source codes with feed-forward 45
2.5.2 'Bad' source codes with feed-forward 47
2.6 Feed-Forward with arbitrary delay 49
2.7 Conclusion 51

3. Directed Information for Channel Coding with Delayed Feedback and Common State Information 53

3.1 Introduction . . . 53
3.2 Channel Capacity with delayed feedback . . . 54
3.2.1 Intuition . . . 57
3.3 Channels with feedback and side-information . . . 60
3.3.1 Intuition . . . 63
3.4 Source-channel separation with feed-forward and feedback . . . 66
3.5 Conclusion . . . 72

4. Evaluating the Rate-Distortion Function of Sources with Feed-forward and the Capacity of Channels with Feedback . . . 73

4.1 Introduction . . . 73
4.2 Evaluating the feed-forward rate-distortion function . . . 74
4.2.1 Markov sources with feed-forward . . . 77
4.3 Evaluating the channel capacity with feedback . . . 78
4.3.1 Markov channels with feedback . . . 82
4.4 Examples . . . 83
4.4.1 Source coding examples . . . 83
4.4.2 Channel coding example . . . 95
4.5 Conclusion . . . 100

5. Multiple Descriptions with Feed-forward . . . 101

5.1 Introduction . . . 101
5.2 Problem statement and main results . . . 105
5.2.1 Feed-forward to only one decoder . . . 105
5.2.2 Feed-forward to both decoders 1 and 2 . . . 108
5.3 Example . . . 108
5.4 Proof of Theorem . . . 111
5.5 Conclusion . . . 116

6. Summary and Future Directions . . . 118

6.1 Summary . . . 118
6.2 Future Directions . . . 119

APPENDICES . . . **123**

BIBLIOGRAPHY . . . **165**

LIST OF FIGURES

Figure

1.1 Source coding with side-information 3

1.2 Time-line: instantaneous observations 4

1.3 Time-line: delayed observations 4

1.4 Channel with unit-delay feedback 7

2.1 Source coding system with feed-forward. 18

2.2 Backward test channel interpretation 23

2.3 Code function for a binary source. 28

2.4 Representation of a source coding scheme with feed-forward. 34

2.5 Source coding system with k-delayed feed-forward. 49

3.1 Channel with k-delayed feedback 54

3.2 Channel with k-delayed feedback and l-delayed side-information 60

3.3 Joint source-channel coding system with feedback and feed-forward. 67

4.1 Markov chain representing the stock value 88

4.2 Markov channel $\{P(Y_i|X_i, Y_{i-1})\}$ 96

5.1 The multiple descriptions problem 102

5.2 The multiple descriptions problem with feed-forward 103

5.3 Comparison of multiple descriptions with and without feed-forward 110

5.4 Codebook cells for decoder 1 112

B.1 The role of a code-function 155

LIST OF TABLES

Table

2.1 Time-line for a feed-forward problem with blocklength $N = 5$ 19

4.1 Distortion $e\left(\hat{x}_i, x_{i-1}, x_i\right)$ 84

4.2 The distribution $P\left(x_i|x_{i-1}, \hat{x}_i\right)$ 86

4.3 The conditional distribution $P(\hat{x}_i|x_{i-1}, x_i)$ 86

4.4 Distortion $e\left(\hat{x}_i, x_{i-1} = j, x_i\right)$ 89

4.5 The distribution $P\left(X_i|x_{i-1}, \hat{x}_i\right)$ 89

4.6 The conditional distribution $P(\hat{X}_i|x_{i-1}, x_i)$ 90

5.1 Time-line of events at encoder and decoder with feed-forward with $k = 1$ 113

LIST OF APPENDICES

Appendix

A. Proofs for Chapter 2 124

A.1 Proof of Lemma 2.4(AEP) 124
A.2 Proof of Direct Part of Theorem 2 128
A.2.1 Proof of Lemma A.7 138
A.3 Proof of Converse Part of Theorem 2 140
A.3.1 Proof of Lemma A.8 142
A.4 Proof of Theorem 3 144
A.5 Proof of Theorems 4 and 5 149
A.6 Proof of (2.69) 151

B. Proofs for Chapter 3 152

B.1 Proof of Lemma 3.3 152
B.2 Proof of Theorem 9 153
B.2.1 The joint distribution 153
B.2.2 Proof of capacity result- Direct part 155
B.2.3 Converse part 157

C. Proofs for Chapter 4 158

C.1 Proof of Theorem 11 158
C.2 Proof of Corollary 4.2 162
C.3 Proof of Theorem 12 163

ABSTRACT

As networked communication systems become increasingly sophisticated, understanding information flow in networks is a problem of central importance. Multi-user information theory attempts to understand various interactions in a network by studying small building blocks such as distributed compression, multiple access, broadcast etc. In this thesis, we investigate two problems in network communication from an information-theoretic perspective: feed-forward in sources and feedback in channels. The two problems are closely related in the sense that there is a dynamic aspect to either the decoder or the encoder in each of them.

Feedback in channels has been extensively studied, but the problem of source coding with feed-forward is recent. In source coding with feed-forward, to reconstruct each source sample, the decoder has knowledge of some past source samples in addition to the codebook index. This extra information can presumably help the decoder produce a better reconstruction. We obtain the optimal rate-distortion function and characterize the error exponent for this problem. The relevant information quantity is directed information, which captures the causal flow of information from one random sequence to another.

Directed information was also recently used to characterize the capacity of channels with feedback. We provide an interpretation of directed information that helps understand why it arises in the context of feedback. This interpretation is used to obtain the feedback capacity of channels with side-information available with delay at both the encoder and decoder.

It turns out that feed-forward/feedback does not improve the performance limit for

memoryless sources/channels. Hence we consider general sources and channels with memory. Consequently, the rate-distortion function and capacity are multi-letter expressions that cannot be computed easily in general. We present an approach that can used to compute these expressions for a wide class of sources, channels and distortion measures.

Finally, we investigate the effect of feed-forward in multiple descriptions coding. We obtain a computable 'single-letter' achievable rate region. In contrast to the point-to-point case, even for i.i.d sources, the obtained rate-region is strictly better than the best known region without feed-forward.

CHAPTER 1

Introduction

Claude Shannon's seminal 1948 paper [73] gave birth to the field of information theory that has been the bedrock for building modern communication and storage systems. The spectacular success and far-reaching impact of the theory so far has largely been in the design of point-to-point communication systems. It can be argued that 'network' information theory (whose roots can be also be traced back to Shannon [75]) has not yet had a similar impact on the design of multi-terminal networks. Networked communication systems have become increasingly sophisticated in the past decade. Two prime examples are high-speed computer and wireless networks designed to deliver rich multimedia content and low-power sensor networks aiming to measure and process large amounts of data with distributed resources. After a relatively quiet period in the late eighties and early nineties, the emergence of these network applications led to renewed interest in multi-user information theory.

It is evident that understanding information flow in networks and developing efficient codes for network communication are problems of central importance. There are a number of features that make the models in multi-user information theory interesting (and challenging): feedback, side-information, distributed compression and transmission are some of these features. This thesis focuses primarily on two of

these elements- feed-forward in sources and feedback in channels. The problems of feed-forward and feedback are closely related and may be considered dual problems of one another.

Channels with feedback have been studied extensively, but the problem of source coding with feed-forward is fairly recent. As is well known, an 'encoder' and a 'decoder' are the two basic components of a communication system. In the channel feedback problem, there is a dynamic aspect to the encoder, while in the source feed-forward problem, there is a dynamic aspect to the decoder. Consequently, a notion of information which captures the dynamics is required to characterize the performance limit of these problems. The appropriate information quantity is the *directed information*, which captures the causal flow of information between random sequences. New tools (such as a generalized notion of typicality, which we introduce) are also required to characterize the interactions in systems with feed-forward/feedback. In the remaining sections of this chapter, we motivate and explain these problems, and outline the contributions of each chapter of the thesis.

Before we proceed, a word about the notation used in the rest of this thesis. Random variables are denoted with capital letters and their realizations with lowercase letters. Superscripts will be used to denote vectors of random variables. For example, A^n will denote the vector $(A_1, \ldots, A_n)$. Bold-face letters will be used for random processes. Thus $\mathbf{A}$ will denote the random process $\{A_n\}_{n=-\infty}^{\infty}$, where A_n is the random variable corresponding to time instant n.

1.1 Source coding with feed-forward

With the recent emergence of applications involving sensor networks [42], the problem of source coding with side-information at the decoder [89] has gained special

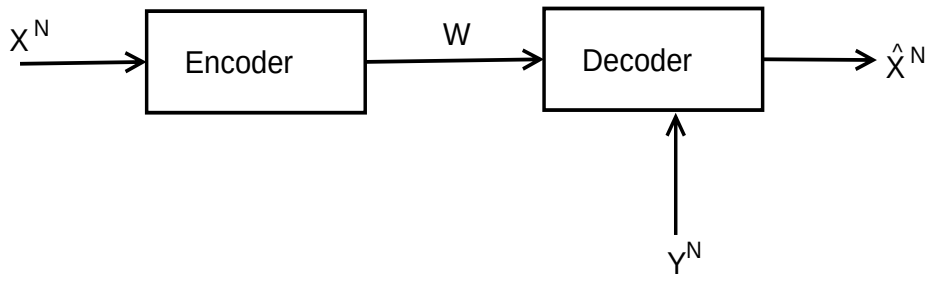

Figure 1.1: Source coding with side-information

significance. The problem is depicted in Figure 1.1. The source of information, modeled as a random process $\mathbf{X} = \{X_n\}_{n=1}^{\infty}$, is encoded in blocks of length N into a message W. W is then transmitted over a noiseless channel of finite rate to a decoder, which has access to some side information $\mathbf{Y} = \{Y_n\}_{n=1}^{\infty}$ that is correlated to the source $\mathbf{X}$. The decoder with the help of the side information $\mathbf{Y}$ and the message W obtains an optimal estimate of N samples of the source at once, and hence, over time, a reconstruction of the process $\mathbf{X}$. The goal is to minimize the reconstruction distortion for a fixed transmission rate. The optimal rate-distortion performance limit for this problem when $(\mathbf{X}, \mathbf{Y})$ is a joint i.i.d process was obtained by Wyner and Ziv in [89]. This problem is used to model the compression problem in general sensor networks where X and Y are the correlated signals captured by the sensor and the destination nodes. The Wyner-Ziv problem has also been the basis for designing for distributed video coding schemes (see [69], for example).

In the problem described above, the encoder and decoder are time-synchronous, i.e., to reconstruct a set of N samples of X, the decoder uses the corresponding set of N samples of Y. The implicit assumption is that the underlying sample pairs (X_i, Y_i) are simultaneously observed at the encoder and the decoder, respectively. So after an encoding delay of N samples, when the decoder gets the message W , it has access to the corresponding N samples of Y, so that the decoding can begin

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	–	–	–	–	W	–	–	–	–	W
Side info	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Figure 1.2: Time-line: instantaneous observations

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	–	–	–	–	W	–	–	–	–	W
Side info	–	–	–	–	–	–	Y_1	Y_2	Y_3	Y_4
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Figure 1.3: Time-line: delayed observations

immediately. The time-line of the samples of the source, the message and the side information is depicted in Figure 1.2 for $N = 5$. Note that in this model, at the 6th time unit, the decoder reconstructs $\hat{X}_1, \ldots \hat{X}_5$ simultaneously as a function of W and $Y_1, \ldots Y_5$, though it may display them as shown in Figure 1.2.

Often the side-information Y is a noisy version of the source X which is assumed to be available simultaneously at the decoder. The question we would like to ask is: *what happens if the underlying signal field X is traveling slowly from the location of the encoder to that of the decoder, where it is available as Y ?* In other words, there is a delay between the instant when ith source sample X_i is observed at the encoder and the instant when corresponding noisy version Y_i is observed at the decoder.

Figure 1.3 shows such a model when the signal field delay is 6 time units and block length $N = 5$. Once the index W is received, we want the reconstruction to be real-time and sequential. In Figure 1.3, for real-time reconstruction of the ith source sample, all the past $i - 1$ samples of the side information are available. In

other words, the decoding operation consists of a sequence of functions such that the ith reconstruction is a function of W and $(i-1)$ side information samples. So we need a new dynamic compression model that takes the important physical signal delay into account in its real-time reconstruction. We refer to this model as source coding with feed-forward. Note that the encoding operation, however, remains as in the Wyner-Ziv problem- a mapping from the N-product source alphabet to an index set of size 2^{NR} where R is the rate of transmission. Thus in this problem, the encoder is non-causal and the decoder is causal.

In Chapter 2, as a first step, we consider an idealized version of this problem where we assume that the traveling source field X is available noiselessly with an arbitrary delay at the decoder, i.e. $Y = X$. We call this problem source coding with noiseless feed-forward. This was first considered by Weissman and Merhav in [85] in the context of competitive prediction. From Figure 1.3, it is clear that the model with $Y = X$ is meaningful only when the delay is at least $N+1$, where the block length is N. However, for a general Y, any delay leads to a valid problem. When the delay is $N+k$, we refer to the problem as *source coding with delay k feed-forward*. Thus with delay k feed-forward, the decoder has available the source samples until time $i-k$ as side-information to reconstruct $\hat{X}_i$. We should mention that source coding with feed-forward can be considered the dual problem of channel coding with feedback, which we discuss in Section 1.2.

1.1.1 Contributions

The main contributions of Chapter 2 can be summarized as follows.

- We first characterize the optimal rate-distortion function for a general discrete source with a general distortion measure and with noiseless delay 1 feed-forward.

This function, denoted $R_{ff}(D)$, is given by the minimum of the directed information function [53] flowing from the reconstruction to the source. From the properties of directed information, it will follow that $R_{ff}(D) \leq R(D)$, where $R(D)$ denotes the optimal Shannon rate-distortion function for the source without feed-forward.

- We extend the Asymptotic Equipartition Property [15] to define a new kind of typicality that we call 'directed typicality'. This is used to provide a simple, intuitive direct coding theorem for stationary, ergodic sources with feed-forward.
- The performance of the best possible source code (with feed-forward) of rate R, distortion D and block length N is asymptotically characterized by an error exponent. We characterize the error exponent for a general source with feed-forward.
- The results are then extended to feed-forward with arbitrary delay k. Using an intuitive interpretation of directed information, we introduce a generalized form of directed information to analyze the problem of source coding with delay-k feed-forward. In particular, the optimal rate-distortion function as well as the error exponent is characterized for delay k feed-forward.

1.2 Channel coding with feedback

Feedback is widely used in modern communication systems to enhance reliability of transmission. Since the early days of information theory, feedback has been an important topic of research. Claude Shannon even chose it as the topic of the inaugural Shannon lecture in 1973. Figure 1.4 depicts a channel with feedback. W is the message to be reconstructed, and the channel input and output at time n are denoted X_n and Y_n, respectively. The non-anticipatory channel is defined as a

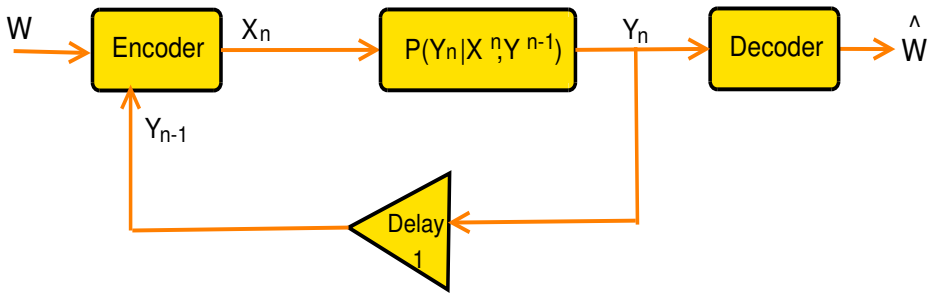

Figure 1.4: Channel with unit-delay feedback

sequence of distributions $\{P(Y_n|X^n, Y^{n-1})\}_{n=1}^{\infty}$. The channel has a feedback delay equal to one time unit- to produce X_n, the decoder has complete knowledge of the first $n-1$ channel outputs Y^{n-1}.

In one of the earliest results concerning feedback, Shannon showed in [74] that feedback does not increase the capacity of a point-to-point discrete memoryless channel (DMC). We recall that a discrete memoryless channel is defined by the property

$$P(Y_n|X^n, Y^{n-1}) = P(Y_n|X_n), \quad \forall n.$$

For point-to-point channels with memory, feedback can indeed increase the capacity. However, for a long time, information theorists lacked the tools to characterize the capacity of these channels. What they lacked was a way to give 'direction' to information flow. Marko was one of the first to address this deficiency in [50]. His aim was to develop a *bidirectional communication theory* to describe the generation and processing of information when living beings, especially humans, interacted with each other. Quoting from his paper [50]:

> The conventional information theory is capable of giving a generalized measure of correlation, but not of distinguishing the direction of information flow. This exactly is the aim of the bidirectional communication theory.

Marko defined new information quantities, that he called 'free information' and 'directed transinformation', to capture the statistical coupling between interacting

systems. Inspired by Marko's work, Massey [53] introduced the concept of directed information flowing from a random sequence X^N to a random sequence Y^N. Massey also pointed out that the prevailing definition of a discrete memoryless channel was incorrect- in the sense that it implicitly prohibited feedback- and gave a correct definition in [53]. Kramer [46, 47] used directed information to characterize the capacity region of the multiple-access channel with noiseless feedback and also the two-way channel. Tatikonda later established the capacity of a general single-user channel with feedback in terms of the directed information flowing from the input to the output [77].

In Section 1.2.1, we outline the contributions of Chapter 3 to the feedback problem. First, we review some important existing results on channel coding with feedback. In a result that complemented the fact that feedback does not increase the capacity of a discrete memoryless channel, Dobrushin showed in [21] that for sufficiently symmetric discrete memoryless channels, noiseless feedback does not improve the block-coding error exponent either. Despite these two negative results, there is a slew of positive results showing feedback does help in many settings. Schalkwijk and Kailath showed in [72] that for AWGN channels with an average power constraint, noiseless feedback enables a doubly exponential decay in probability or error with increasing block-length. If we do not restrict ourselves to block codes and consider variable delay coding instead, noiseless feedback can increase the error exponent as shown in various important papers [38, 25, 10, 90].

In contrast to the single-user case, feedback can enlarge the capacity region of discrete memoryless multi-user channels. This was first demonstrated in the case of a multiple access channel (MAC) by Gaarder and Wolf [26]. An achievable rate-region for a discrete memoryless MAC was determined by Cover and Leung [17],

which was recently improved upon [9]. Willems [86] showed that that the Cover-Leung region is the capacity region for a certain class of MACs with feedback. The feedback capacity region of the Gaussian MAC, which does not belong to this class, was determined by Ozarow in [59]. Feedback can also increase the capacity region of a discrete memoryless broadcast channel as shown by example in [23]. It should be noted though, that feedback does not increase the capacity of a physically degraded broadcast channel [28]. The literature on feedback is vast and the list above is by no means exhaustive. We shall mention some other important works in Chapter 3.

1.2.1 Contributions

The contributions of Chapter 3 can be summarized as follows.

- We present an interpretation of directed information that yields some insight about its relevance in the context of feedback capacity. This intuition is then used to characterize the capacity of channels with delayed feedback and channel state information which is available with some delay at both encoder and decoder.
- We consider a system where a source has to be transmitted over a channel with delay l feedback. The source is reconstructed with delay k feed-forward to a specified distortion level. Under suitable conditions of information stability, a source-channel separation theorem is established.

1.3 Evaluating the expressions

One of the most appealing features of Shannon's formulas for channel capacity and the rate-distortion function is the simplicity of the optimizations involved. Recall that the capacity C of a discrete memoryless channel $P_{Y|X}$ (with input X and output

Y) is given by

$$C = \max_{P_X} I(X;Y)$$

and the rate-distortion function $R(D)$ for a source P_X with reconstruction $\hat{X}$ is given by

$$R(D) = \min_{P_{X|\hat{X}}: E[d(X;\hat{X})] \leq D} I(X;\hat{X}).$$

These are 'single-letter' optimizations, which means that we need to optimize over probability distributions of a single random variable. Efficient techniques such as the Blahut-Arimoto algorithm [7] exist to compute these single-letter optimizations.

For problems with feed-forward and feedback, there is no improvement in the performance limit when the source/channel is memoryless; the interesting sources and channels are those with memory. Further, due to the dynamics introduced by feed-forward and feedback, we cannot expect the optimal joint distributions to be stationary and ergodic in general. Consequently, the formulas (involving directed information, cf. Chapters 2 and 3) for the optimal rate-distortion function with feed-forward and channel capacity with feedback are multi-letter expressions. This means that we cannot hope to have a simple Blahut-Arimoto like algorithm to perform the optimizations.

Computing the performance limits with feed-forward/feedback is an important problem. In Chapter 4, we take a different approach to the problem of computing the rate-distortion and capacity expressions (with feed-forward and feedback, resp.). We obtain the structure of the distortion (cost, resp.) function in order for a given joint distribution to achieve the optimum rate-distortion function (channel capacity, resp.). For discrete memoryless channels and sources without feedback/feed-forward, such an approach appears in the book of Csiszár and Körner [19, p. 147, Problems 2,3] to characterize the structure of the cost/distortion function. Our results may be

viewed as an extension of the results in [19] to problems with delayed feedback and feed-forward. This approach is especially relevant since it is otherwise infeasible to calculate the performance limits with feedback and feed-forward.

1.3.1 Contributions

For the feed-forward and feedback problems, we derive structural results that relate the structure of the distortion/cost function to that of the optimal joint distribution. In particular,

- The structure of the distortion function for a chosen joint process (satisfying suitable conditions) to achieve the optimum rate-distortion function with feed-forward is characterized.
- The structure of the cost function for a chosen joint process (satisfying suitable conditions) to achieve the capacity of a feedback channel is characterized.
- Examples are provided to illustrate how the above results can be used to compute the feed-forward rate-distortion function and feedback capacity for many sources and channels.

Our structural results are established under the conditions that the joint process is information stable. The definition of information stability is found in Chapter 4- this condition encompasses a wide range of ergodic processes including those that are stationary as well as asymptotically stationary.

1.4 Multiple Descriptions with feed-forward

In Chapter 5, we investigate the role of feed-forward in a multi-terminal problem, viz. multiple descriptions source coding . The multiple descriptions problem was introduced by Gersho, Witsenhausen, Wyner, Ziv and Ozarow in the 1979 IEEE

Information Theory Workshop. As a motivating example, consider a packet-switched network in which a source node compresses data into packets, which are then sent to a destination node. There is a chance that packets may be dropped in the network. To ensure reliable transmission, each source sequence can be compressed to two different packets that are both sent to the destination. If either packet is received, the source data is reconstructed with adequate quality, but we would like better reconstruction quality if both packets are received. Thus, in this situation, the two individual reconstructions of the same source sequence need to refine each other. Given certain levels of distortion to be achieved (depending on the number of packets received), the goal of the multiple descriptions problem is to determine the set of achievable rates of compression for each of the individual packets. An achievable rate region for this problem for an i.i.d source was first determined by El Gamal and Cover [16], which was improved by Zhang and Berger[92]. The optimal rate region for multiple descriptions of an i.i.d source is still an open problem and known only for certain special cases (cf. [1, 58]).

Our aim in Chapter 5 is to explore how (partial or complete) feed-forward can help in this multi-terminal setting. The following problem is another example that motivates our study of multiple descriptions with feed-forward. There are four people named Alice, Bob, Carol and Dave. Alice has an equiprobable binary source that Bob, Carol and Dave are interested in reconstructing. Bob and Carol each want to reconstruct with the fraction of their errors being at most d, while Dave needs error-free reconstruction. Bob and Carol agree to buy some information from Alice separately, and Dave agrees to buy the information available to both Bob and Carol. Further assume that after reconstruction of each source sample, Alice reveals to Carol (but not Bob and Dave) the actual value of the sample. The minimum rates

of information that Alice would have to supply to Bob and Carol in this situation is the multiple description rate-distortion region with feed-forward to Carol only.

1.4.1 Contributions

The contributions in Chapter 5 are listed below.

- The multiple descriptions problem for an i.i.d source with feed-forward to one of the side decoders is considered. We obtain a single letter achievable rate region for this setting. The rate-region is obtained using a coding strategy that uses feed-forward to build correlation between the two reconstructions cheaply. We recall that in the multiple descriptions problm, the individual reconstructions need to be correlated in order to refine one other.
- We show that the achievable rate-region obtained (for feed-forward to only one of the decoders) is larger than the Zhang-Berger region – the best known rate-region without feed-forward.
- A lower bound on the minimum sum-rate required by Bob and Carol in the above example was obtained in [92]. We evaluate our rate-region (with feed-forward to only Carol) and show that a sum-rate smaller than the lower bound can be achieved. In other words, for this i.i.d multiple descriptions problem, feed-forward to just one decoder enables rates better than the optimal rates without feed-forward.

We observe that feed-forward can strictly improve the optimal rate-region even for i.i.d source. This is in contrast to point-to-point source coding, where feed-forward improves the rate-distortion function only for sources with memory.

CHAPTER 2

Source Coding with Feed-Forward

2.1 Introduction

The problem of source coding with feed-forward was motivated in the previous chapter (cf. Section 1.1) by an application that involved compressing samples of a random field at one node of a network, and reconstructing them sequentially at a different location. However, the relevance of source coding with feed-forward extends much beyond the sensor networks application outlined above. It is also closely related to the problem of prediction. In fact, it was first considered in the context of competitive prediction in [85].

As an example, suppose that we want to predict the stock price of some company over an N-day period. Let the share price on day n be X_n. At the beginning of day n, we have to make our prediction $\hat{X}_n$ for the closing price of the share on that day. Let $d(X_n\hat{X}_n)$ be a measure of our guessing error for day n. We want to minimize our average error over an N-day period:

$$\frac{1}{N}\sum_{n=1}^{N} d(x_n, \hat{x}_n).$$

Note that at the end of each day, we know the actual closing price of the stock for that day. Hence, at the time we predict $\hat{X}_n$, we know X^{n-1}, the actual values of

the stock price on the first n days [1].

Further suppose that there is an insider who has *a priori* knowledge of how the stock is going to behave over the N days. At the beginning of the N-day period, she is willing to sell information at some finite rate R bits/day to aid our prediction Over the N-day period, if we want our average prediction error to satisfy

$$\frac{1}{N}\sum_{n=1}^{N} d(x_n, \hat{x}_n) \leq D, \tag{2.1}$$

what is the minimum rate R needed? This problem of 'prediction with a priori information' is identical to source coding with feed-forward. In Chapter 4, we will discuss this problem in detail for a Markov source and Markovian error function.

The problem of source coding with noiseless feed-forward was first considered by Weissman and Merhav in the context of competitive prediction in [85]. They consider sources with feed-forward delay 1 and a single-letter, difference distortion measure. In [85], the optimal distortion-rate function with feed-forward is derived for sources that can be represented auto-regressively with an innovations process that is either IID or satisfies the Shannon Lower Bound (SLB)[15] with equality. The distortion-rate function was evaluated in [85] for a symmetric binary Markov source with feed-forward and a stationary Gaussian source with feed-forward as examples of this result. For sources with general innovations processes, [85] provides upper and lower bounds on the distortion-rate function. The block coding error exponent is also derived in [85] for the case where the innovations process is IID and is shown to be the same as Marton's no-feed-forward error exponent [52]. It was noted in [85] that feed-forward can only decrease the distortion-rate function of a source; however, for IID sources and all sources that satisfy SLB with equality and with

[1] We will use the superscript notation to denote a sequence of random variables. Thus $X^{n-1} = [X_1, \ldots, X_{n-1}]$.

single-letter difference distortion measures, feed-forward does not reduce the optimal distortion-rate function.

Later, the model of source coding with general feed-forward was considered in [64] as a variant of the problem of source coding with side information at the decoder, and a quantization scheme with linear processing for IID Gaussian sources with mean squared error distortion function and with noiseless feed-forward was reported. It was also shown that this scheme approaches the optimal rate-distortion function. In [51], an elegant variable-length coding strategy to achieve the optimal Shannon rate-distortion bound for any finite-alphabet IID source with feed-forward was presented, along with an illustrative example. The problem of source coding with feed-forward is also related to source coding with a delay-dependent distortion function[48], causal source coding[56] and real-time source coding[88] .

The main results of this chapter can be summarized as follows:

1. The optimal rate-distortion function for a general discrete source with a general distortion measure and with noiseless feed-forward, $R_{ff}(D)$, is given by the minimum of the directed information function [53] flowing from the reconstruction to the source. From the properties of directed information, it will follow that $R_{ff}(D) \leq R(D)$, where $R(D)$ denotes the optimal Shannon rate-distortion function for the source without feed-forward.

2. We extend the Asymptotic Equipartition Property [15] to define a new kind of typicality that we call 'directed typicality'. This is used to provide a simple, intuitive direct coding theorem for stationary, ergodic sources with feed-forward.

3. The performance of the best possible source code (with feed-forward) of rate R, distortion D and block length N is asymptotically characterized by an error exponent. We characterize the error exponent for a general source with feed-

forward.

4. Extension of these results to feed-forward with arbitrary delay. We introduce a generalized form of directed information to analyze the problem of source coding with delayed feed-forward.

We now briefly outline how our results differ from that of [85]. In [85], feed-forward is considered in the context of competitive prediction. The optimal distortion-rate function of a source with feed-forward is completely characterized in [85] only when the source has an autoregressive representation with an innovations process that is either IID or satisfies the Shannon Lower Bound with equality. This characterization of the distortion-rate function is in terms of the innovations process. In our work, we derive the optimal rate-distortion function with feed-forward for any general source with feed-forward. This is expressed in terms of directed information, a quantity involving just the source X and the reconstruction $\hat{X}$. The feed-forward error exponent is derived in [85] for sources with an autoregressive representation with IID innovations. We characterize the error exponent for a general source. The results in [85] are derived for single-letter, difference distortion measures and feed-forward with delay 1. Our results are derived for arbitrary (not necessarily single-letter) distortion measures and feed-forward with arbitrary delay.

This chapter is organized as follows. In Section 2.2 we give a fairly formal definition of the above source coding model and the intuition behind the proposed approach. Instead of giving the main result for the most general sources and then considering the special cases, we first consider the special case when the source and the reconstruction processes are jointly stationary and ergodic and give a direct coding theorem in Section 2.3 which captures the essence of this problem. We must mention here that for stationary, ergodic sources *without* feed-forward with single-

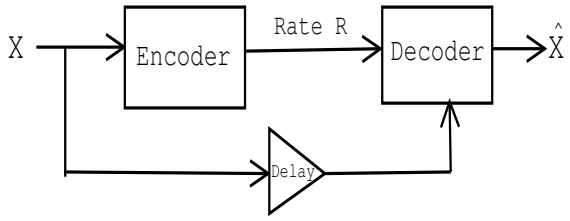

Figure 2.1: Source coding system with feed-forward.

letter distortion measures, the optimal rate-distortion function is attained by a jointly stationary and ergodic $(X, \hat{X})$ process [32]. Unfortunately, a similar result may not hold for stationary, ergodic sources with feed-forward even with single-letter distortion measures. This is because the information available at the decoder changes with time. Hence we can only obtain a direct coding theorem by restricting our attention to stationary, ergodic joint processes.

To obtain a tight rate-distortion theorem, we have to consider general processes. The method of information spectrum introduced by Han and Verdu [33] is a powerful tool to deal with general processes. Using this, we give the direct and converse coding theorems for general sources in Section 2.4. In that section we also consider some special cases such as discrete memoryless sources and Gaussian sources. Error exponents are considered in the general setting in Section 2.5. We extend our results to arbitrary delays in Section 2.6 and finally, concluding remarks are given in Section 2.7.

2.2 The source coding model

2.2.1 Problem statement

The model is shown in Figure 2.1. Consider a general discrete source X with Nth order probability distribution P_{X^N}, alphabet $\mathcal{X}$ and reconstruction alphabet $\widehat{\mathcal{X}}$. There is an associated distortion measure $d_N : \mathcal{X}^N \times \widehat{\mathcal{X}}^N \to \mathbb{R}^+$ on pairs of sequences. It is assumed that $d_N(x^N, \hat{x}^N)$ is normalized with respect to N and is

Table 2.1: Time-line for a feed-forward problem with blocklength $N = 5$

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	-	-	-	-	W	-	-	-	-	W
Extra info	-	-	-	-	-	-	X_1	X_2	X_3	X_4
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

uniformly bounded in N.

Definition 2.1. An $(N, 2^{NR})$ source code with feed-forward of block length N and rate R consists of an encoder mapping e and a sequence of decoder mappings $g_n, n = 1, \ldots, N$:

$$e : \mathcal{X}^N \to \{1, \ldots, 2^{NR}\}$$

$$g_n : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{n-1} \to \widehat{\mathcal{X}}, \quad n = 1, \ldots, N$$

The encoder and decoder mappings are

$$e(X^N) = W, \qquad g_n(W, X^{n-1}) = \hat{X}_n, \; n = 1, \ldots N$$

where $W \in \{1, \ldots, 2^{NR}\}$ denotes the index and $\hat{X}_n$ is the reconstruction of the nth sample.

The encoder maps each N-length source sequence to an index in $\{1, \ldots, 2^{NR}\}$. Table 2.1 shows the time-line of events when the block-length $N = 5$. As shown in the table, the decoder receives the index W transmitted by the encoder at the end of time instant 5. At time instant 6, $\hat{X}_1$ is reconstructed by the decoder using W. To reconstruct $\hat{X}_2$ at time 7, the decoder knows X_1 in addition to the index. Thus to produce $\hat{X}_n$ (for $2 \leq n \leq 5$), the decoder knows all the past $(n-1)$ samples of the source (in addition to the index W). At the end of time instant 10, the encoder

produces the index corresponding to the next block of source samples $X_6 - X_{10}$, and the decoding of this block in a manner identical to the first block.

Let $\hat{x}^N$ denote the reconstruction of the source sequence x^N. We want to minimize R for a given distortion constraint. We consider two types of distortion constraints: 1) expected distortion constraint and 2) probability-1 distortion constraint. These constraints are formally defined in the sequel. For any D, let $R_{ff}(D)$ denote the infimum of R over all encoder-decoder pairs for any block length N such that the distortion is less than or equal to D. It is worthwhile noting that source coding with feed-forward can be considered the dual problem [66, 4, 13] of channel coding with feedback.

2.2.2 Intuition behind the proposed approach

To analyze the problem of source coding with feed-forward we need a directional notion of information. This is given by directed information, as defined by Massey [53]. This notion was motivated by the work of Marko[50] and was also studied in [30, 11, 71] in the context of dependence and feedback between random processes. More recently, directed information has been used to characterize the capacity of channels with feedback [46, 77].

Definition 2.2. [53] The directed information flowing from a random vector A^N to another random vector B^N is defined as

$$I(A^N \to B^N) = \sum_{n=1}^{N} I(A^n; B_n | B^{n-1}). \tag{2.2}$$

Note that the definition is similar to that of mutual information $I(A^N; B^N)$ except that the mutual information has A^n instead of A^N in the summation on the right. The directed information has a nice interpretation in the context of our problem. We can write the directed information flowing from the reconstruction $\hat{X}^N$ to the source

X^N as

$$I(\hat{X}^N \to X^N) = I(X^N; \hat{X}^N) - \sum_{n=2}^{N} I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}). \tag{2.3}$$

(2.3) can be derived using the chain rule as follows [54].

$$\begin{aligned} I(\hat{X}^N \to X^N) + \sum_{n=2}^{N} I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}) &= \sum_{n=1}^{N} I(\hat{X}^n; X_n | X^{n-1}) + \sum_{n=2}^{N} I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}) \\ &= H(X^N) + H(\hat{X}^N) - \sum_{n=1}^{N} H(X_n | X^{n-1}, \hat{X}^n) - H(\hat{X}_n | X^{n-1}, \hat{X}^{n-1}) \\ &= H(X^N) + H(\hat{X}^N) - \sum_{n=1}^{N} H(X_n, \hat{X}_n | X^{n-1}, \hat{X}^{n-1}) \\ &= H(X^N) + H(\hat{X}^N) - H(X^N, \hat{X}^N) = I(X^N; \hat{X}^N). \end{aligned}$$

Consider first the standard source coding problem without feed-forward. Here the goal is to construct a codebook of length N sequences $\hat{X}^N$ such that for every source sequence X^N, there is at least one sequence in the codebook that is jointly typical with a specified joint distribution $P(X^N, \hat{X}^N)$ (that satisfies some distortion constraint). We know from rate-distortion theory that the minimum rate of such a codebook is $\frac{I(\hat{X}^N; X^N)}{N}$. With feed-forward, since the decoder knows the symbols X^{n-1} to reconstruct $\hat{X}_n$, (2.3) says we need not spend $I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1})$ bits to code this information. Loosely speaking, this rate comes for 'free' because of feed-forward. In other words, the performance limit on this problem is given by the minimum of the directed information.

An interesting way to understand any source compression system is to analyze the corresponding backward test channel [5, 15, 63]. This is a fictitious channel which connects the source with the reconstruction, characterized by the conditional distribution of the source given the reconstruction. In source coding with feed-forward, the decoder first gets the index W (sent by the encoder) containing the information about the first N samples of X. The process of reconstruction starts

with the reconstruction of the first sample $\hat{X}_1 = g_1(W)$ as a function of W alone. In the next clock cycle, the decoder has W and X_1. This can be interpreted as follows: $\hat{X}_1$ goes through a non-anticipatory fictitious channel to produce X_1 and is fed back to the decoder. Now the decoder reconstructs the second sample $\hat{X}_2 = g_2(W, X_1)$ as a function of W and X_1. In the next clock cycle, it gets X_2. As before, we can interpret it as $\hat{X}_2$ going through the test channel to produce X_2 which is fed back to the decoder and so on. So this test channel can be thought of as having $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N$ as input and $X_1, X_2, \ldots, X_N$ as output with a sequence of conditional distributions given by

$$\hat{Q}_1(X_1|\hat{X}_1), \hat{Q}_2(X_2|X_1, \hat{X}_1, \hat{X}_2), \ldots, \hat{Q}_i(X_i|X^{i-1}, \hat{X}^i), \ldots, \hat{Q}_N(X_N|X^{N-1}, \hat{X}^N),$$

where X^i denotes the vector of $X_1, X_2, \ldots, X_i$. This sequence of conditional distributions is related to the source and the encoder transformation in the following way. Note that the source distribution $P_{X^N}(X^N)$ and the quantizer transformation $P_{\hat{X}^N|X^N}(\hat{X}^N|X^N)$ fix the joint distribution $P_{X^N,\hat{X}^N}(X^N, \hat{X}^N)$. This can be factored into two components as follows:

$$P_{X^N,\hat{X}^N}(X^N, \hat{X}^N) = \prod_{i=1}^{N} P_i(X_i, \hat{X}_i|X^{i-1}\hat{X}^{i-1}) = \prod_{i=1}^{N} Q_i(\hat{X}_i|X^{i-1}, \hat{X}^{i-1}) \prod_{i=1}^{N} \hat{Q}_i(X_i|X^{i-1}\hat{X}^i),$$

where Q characterizes the decoder reconstruction function, whereas $\hat{Q}$ denotes the test channel conditional distribution, and both of them are assumed to have memory. This is illustrated in Fig. 2.2.

2.3 Stationary and ergodic joint processes

In this section, we will provide a direct coding theorem for a general source with feed-forward assuming that the joint random process $\{X_n, \hat{X}_n\}$ is discrete, stationary and ergodic [31]. This assumption is not necessary to prove the rate-distortion theorem for arbitrary sources with feed-forward in Section 2.4 - the purpose is to

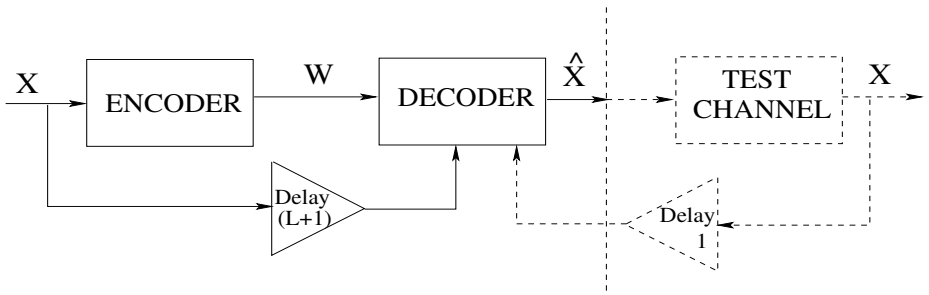

Figure 2.2: Backward test channel interpretation

first give intuition about how feed-forward helps in source coding. This assumption of stationarity and ergodicity leads to a rather simple and intuitive proof of the rate-distortion theorem along the lines of the proof of the rate-distortion theorem for discrete memoryless sources in [15]. We will use a new kind of typicality, tailored for our problem of source coding with feed-forward. A word about the notation before we state the theorem. All logarithms used in the sequel are assumed to be with base 2, unless otherwise stated. The source distribution, defined by a sequence of finite-dimensional distributions [33] is denoted by

$$\mathbf{P_X} \triangleq \{P_{X^n}\}_{n=1}^{\infty}. \tag{2.4}$$

Similarly, a conditional distribution is denoted by

$$\mathbf{P_{\hat{X}|X}} \triangleq \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}. \tag{2.5}$$

Finally, for stationary and ergodic joint processes, the directed information rate exists and is defined by [46]

$$I(\hat{X} \to X) = \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N). \tag{2.6}$$

We use an expected distortion criterion here. For simplicity (only for this section), we assume $d_N(x^N, \hat{x}^N) = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_i)$, where $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. Let d_{max} be the maximum of $d(x, \hat{x}) \quad \forall x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$. Since the distortion measure is bounded, $\lim_{N\to\infty} E[d_N(X^N, \hat{X}^N)]$ exists.

Definition 2.3. R is an achievable rate at expected distortion D if $\forall \epsilon > 0$, for all sufficiently large N, there exists an $(N, 2^{NR})$ code such that

$$E_{X^N}\left[d_N(X^N, \hat{X}^N)\right] \leq D + \epsilon,$$

where $\hat{X}^N$ denotes the reconstruction of X^N.

Theorem 1. *For a discrete stationary and ergodic source X characterized by a distribution $\mathbf{P_X}$, all rates R such that*

$$R \geq R^*(D) \triangleq \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lim_{N\to\infty} E[d_N(X^N,\hat{X}^N)]\leq D} I(\hat{X} \to X)$$

are achievable[2] *at expected distortion D.*

Proof. We first lay down the necessary definitions and lemmas required for the proof. Let $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ be any conditional distribution such that the resulting joint process $\mathbf{P}_{\mathbf{X},\hat{\mathbf{X}}}$ is stationary and ergodic. Since the AEP holds for discrete, stationary and ergodic processes [15], we have

$$\begin{aligned} -\frac{1}{N}\log P(X^N) &\to H(X) \quad \text{w.pr.1}, \\ -\frac{1}{N}\log P(X^N, \hat{X}^N) &\to H(X, \hat{X}) \quad \text{w.pr.1}, \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} H(X) &= \lim_{N\to\infty} H(X_N|X^{N-1}) = \lim_{N\to\infty} \frac{1}{N} H(X^N), \\ H(X, \hat{X}) &= \lim_{N\to\infty} H(X_N, \hat{X}_N|X^{N-1}, \hat{X}^{N-1}) = \lim_{N\to\infty} \frac{1}{N} H(X^N, \hat{X}^N). \end{aligned}$$

We now define two 'directed' quantities, introduced in [50] and [47], respectively. These were used in [77] in the context of channels with feedback. These will be frequently used in the rest of this chapter. $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$,

$$\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \triangleq \prod_{i=1}^{N} P_{\hat{X}_i|\hat{X}^{i-1},X^{i-1}}(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}), \tag{2.8}$$

[2]The infimization is over all conditional distributions such that the joint process $(\mathbf{X}, \hat{\mathbf{X}})$ is stationary and ergodic.

$$\vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \triangleq \prod_{i=1}^{N} P_{X_i|\hat{X}^i,X^{i-1}}(x_i|\hat{x}^i, x^{i-1}). \tag{2.9}$$

These can be pictured in terms of the backward test channel from $\hat{X}$ to X. (2.8) describes the sequence of input distributions to this test channel and (2.9) specifies the backward test channel. Recall that the joint distribution can be split as

$$P_{X^N,\hat{X}^N}(x^N, \hat{x}^N) = \prod_{i=1}^{N} P_{\hat{X}_i|\hat{X}^{i-1},X^{i-1}}(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) \cdot P_{X_i|\hat{X}^i,X^{i-1}}(x_i|\hat{x}^i, x^{i-1}). \tag{2.10}$$

The basic ingredient in our proof is the following Lemma which says that a property analogous to the AEP holds for the directed quantities defined in (2.8) and (2.9). Let

$$H(\hat{X}^N||X^N) \triangleq \sum_{i=1}^{N} H(\hat{X}_i|\hat{X}^{i-1}, X^i).$$

$H(\hat{X}^N||X^N)$ is known as the entropy of $\hat{X}^N$ causally conditioned on X^N [46, 54]. We will also use $H(\hat{X}^N||0X^{N-1})$, the entropy of $\hat{X^N}$ causally conditioned on the delayed X sequence $0X^{N-1}$, which is shorthand for $(0, X_1, X_2, \ldots, X_{N-1})$

Lemma 2.4. *If the process $\{X_i, \hat{X}_i\}_{i=1}^{\infty}$ is stationary and ergodic, we have*

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) \to \vec{H}(\hat{X}||X) \quad w.p.1, \tag{2.11}$$

where

$$\begin{aligned} \vec{H}(\hat{X}||X) &\triangleq \lim_{N\to\infty} \frac{1}{N} H(\hat{X}^N||0X^{N-1}) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} H(\hat{X}_i|X^{i-1}, \hat{X}^{i-1}) \\ &= \lim_{N\to\infty} H(\hat{X}_N|X^{N-1}, \hat{X}^{N-1}), \end{aligned} \tag{2.12}$$

where $0X^{N-1}$ denotes the sequence $[\_, X_1, X_2, \ldots, X_{N-1}]$.

The proof of the lemma is similar to the Shannon-McMillan-Breiman Theorem in [15, 2] and is given in Appendix A.1. We now define a new kind of joint distortion

typicality. Given the source $\mathbf{P_X}$, fix any conditional distribution $\mathbf{P_{\hat{X}|X}}$ to get a joint distribution $\mathbf{P_{X,\hat{X}}} = \{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$. Also recall that the distortion is given by $d_N(x^N, \hat{x}^N) = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_i)$.

Definition 2.5. An ordered sequence pair $(x^N, \hat{x}^N)$ with $x^N \in \mathcal{X}^N$ and $\hat{x}^N \in \hat{\mathcal{X}}^N$ is said to be directed distortion ϵ-typical if:

$$\begin{aligned}
\left|-\frac{1}{N}\log P_{X^N}(x^N) - H(X)\right| &< \epsilon \\
\left|-\frac{1}{N}\log P_{X^N,\hat{X}^N}(x^N, \hat{x}^N) - H(X,\hat{X})\right| &< \epsilon \\
\left|-\frac{1}{N}\log \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) - \vec{H}(\hat{X}||X)\right| &< \epsilon \\
\left|d_N(x^N, \hat{x}^N) - Ed_N(X^N, \hat{X}^N)\right| &< \epsilon
\end{aligned}$$

We denote the set of directed distortion ϵ-typical pairs by $\mathcal{A}_\epsilon^N$.

Lemma 2.6. *If an ordered pair $(X^N, \hat{X}^N)$ is drawn from $P_{X^N,\hat{X}^N}$, then*

$$\Pr((X^N, \hat{X}^N) \in \mathcal{A}_\epsilon^N) \to 1 \quad as \quad N \to \infty. \tag{2.13}$$

Proof. From the AEP for stationary and ergodic processes, the first, second and fourth conditions in Definition 2.5 are satisfied with probability 1 as $N \to \infty$. From Lemma 2.4, the third condition is satisfied with probability 1 as $N \to \infty$, proving the lemma. □

Lemma 2.7. *For all $(x^N, \hat{x}^N) \in \mathcal{A}_\epsilon^N$,*

$$\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \geq P_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \cdot 2^{-N(I(\hat{X}\to X)+3\epsilon)}. \tag{2.14}$$

Proof.

$$
\begin{aligned}
P_{\hat{X}^N|X^N}(\hat{x}^N|x^N) &= \frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{P_{X^N}(x^N)} \\
&= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot P_{X^N}(x^N)} \\
&\leq \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot\frac{2^{-N(H(X,\hat{X})-\epsilon)}}{2^{-N(\vec{H}(\hat{X}||X)+\epsilon)}\cdot 2^{-N(H(X)+\epsilon)}} \\
&= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot 2^{N(I(\hat{X}\to X)+3\epsilon)},
\end{aligned}
\tag{2.15}
$$

from which the lemma follows. The last equality in (2.15) can be proved as follows.

(2.16)

$$
\begin{aligned}
H(\hat{X}^N||0X^{N-1}) + H(X^N) - H(X^N,\hat{X}^N) &= H(\hat{X}^N||0X^{N-1}) - H(\hat{X}^N|X^N) \\
&= H(\hat{X}^N) - H(\hat{X}^N|X^N) - [H(\hat{X}^N) - H(\hat{X}^N||0X^{N-1})] \\
&\overset{(a)}{=} I(X^N;\hat{X}^N) - I(0X^{N-1}\to\hat{X}^N) \\
&\overset{(b)}{=} I(\hat{X}^N\to X^N),
\end{aligned}
$$

where (a) follows by writing the definition of directed information in (2.2) in terms of entropies and (b) follows from (2.3). Dividing by N and taking limits we get the result. □

Having established the required definitions and lemmas, we are ready to describe the coding scheme.

Codetrees: In source coding with feed-forward, to produce the ith reconstruction symbol $\hat{x}_i$, the decoder knows the first $i-1$ source samples x^{i-1}. This means that we could have a different reconstruction $\hat{x}_i$ for each x^{i-1}. To capture this, we need the concept of a code-tree, constructed as follows. Let the first input symbol be $\hat{x}_1$. To reconstruct the next symbol $\hat{x}_2$, the decoder knows x_1. Therefore we have $|\mathcal{X}|$ choices for the $\hat{x}_2$ depending on the x_1 observed. For each value of $\hat{x}_2$, we have $|\mathcal{X}|$ choices for $\hat{x}_3$ and so on, thus forming a tree. A code-tree for a system with binary source

and reconstruction alphabets is shown in Figure 2.3. A rate R source code with feed-forward can be visualized in terms of a set (or codebook) of 2^{NR} code-trees. For each value of the index $W \in \{1, \ldots, 2^{NR}\}$, there is a code-tree with paths indexed by $x_1, x_2, \ldots, x_{N-1}$ such that for each n, the reconstruction $\hat{x}_n$ is the path led to by $x_1, \ldots, x_{n-1}$.

The decoder receives the index of the code-tree chosen by the encoder, traces the path along the code-tree using the fed-forward source symbols and produces the reconstruction. For instance, suppose the code-tree in Figure 2.3 is used and the fed-forward sequence, x^{N-1}, is the all zero sequence. The decoder traces the upper-most path on the tree and obtains the reconstruction symbols along that path.

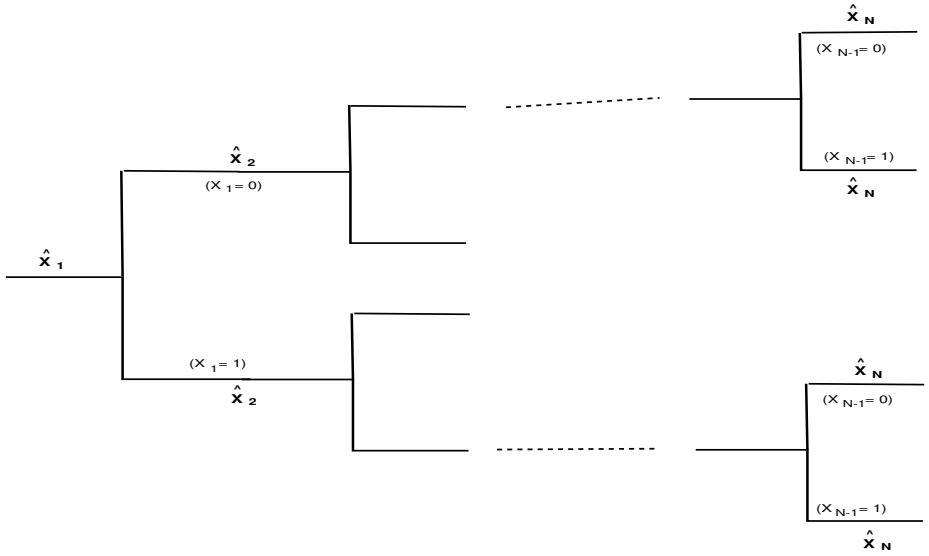

Figure 2.3: Code function for a binary source.

Random generation of code-trees: Pick a joint distribution $\mathbf{P}_{\mathbf{X},\hat{\mathbf{X}}} = \{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$, such that the $X-$marginal has the distribution $\mathbf{P}_{\mathbf{X}}$ and $\lim_{N\to\infty} Ed_N(X^N, \hat{X}^N) \leq D$. This joint distribution is stationary and ergodic by assumption. Fix ϵ and the block length N. Each code-tree is constructed as follows. Pick the first input symbol of the code-tree $\hat{x}_1$ randomly according to the distribution $P_{\hat{X}_1}$. To choose the next symbol,

we have $|\mathcal{X}|$ choices for the $\hat{x}_2$ depending on the x_1 observed. For each possible x_1, $\hat{x}_2$ is chosen randomly and independently according to the distribution $P_{\hat{X}_2|\hat{x}_1,x_1}$ for each possible x_1. For each of these $\hat{x}_2$, there are $|\mathcal{X}|$ possible $\hat{x}_3$'s (depending on the x_2 observed) picked randomly and independently according to the distribution $P_{\hat{X}_3|\hat{x}^2,x^2}$. We continue picking the input symbols in this manner and finally we pick $\hat{x}_N$ according to $P_{\hat{X}_N|\hat{x}^{N-1},x^{N-1}}$. We pick 2^{NR} such code-trees independently in the same fashion to form the codebook of code-trees.

Encoding: We will use jointly typical encoding. The encoder has the sequence x^N. It traces the path determined by x^{N-1} on each of the 2^{NR} code-trees of the codebook. Each of these paths corresponds to a reconstruction sequence $\hat{x}^N[i] \quad (i \in \{1, \ldots 2^{NR}\})$. The encoder chooses a $\hat{x}^N[W]$ that is directed distortion ϵ-typical with x^N and sends W to the decoder. If no such typical $\hat{x}^N$ is found, an encoding error is declared.

Decoding: The decoder receives the index W from the encoder ($W \in \{1, \ldots, 2^{NR}\}$). It uses the Wth code-tree and obtains the reconstruction symbols along the path traced by $\{x_k\}_{k=1}^{N-1}$ that are fed-forward.

Distortion: There are two types of source sequences x^N- a) Good sequences x^N, that are properly encoded with distortion $\leq D+\epsilon$, b) Bad source sequences x^N, for which the encoder cannot find a distortion-typical path. Let P_e denote the probability of the set of bad source sequences for the code. The expected distortion for the code can be written as

$$E[d_N(X^N, \hat{X}^N)] \leq D + \epsilon + P_e d_{max}. \tag{2.17}$$

We calculate the expected distortion averaged over all random codebooks. This is

given by

$$E_{\mathcal{C}}[E[d_N(X^N, \hat{X}^N)]] \leq D + \epsilon + \overline{P}_e d_{max}, \tag{2.18}$$

where $\overline{P}_e$ is the expected probability of the set of bad X^N sequences, the expectation being computed over all randomly chosen codes. We will show that when R satisfies the condition given by Theorem 1, $\overline{P}_e$ goes to 0 as $N \to \infty$. This would prove the existence of at least one rate-R code with expected distortion $\leq D + \epsilon$.

Average Probabilty of Error $\overline{P}_e$: $\overline{P}_e$ is the probability that for a random code $\mathcal{C}$ and a random source sequence X^N, none of the 2^{NR} codewords are jointly typical with X^N. Let $J(\mathcal{C})$ denote the set of good (properly encoded) source sequences for code $\mathcal{C}$. Now,

$$\overline{P}_e = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \sum_{x^N : x^N \notin J(\mathcal{C})} P(x^N) \tag{2.19}$$

$$= \sum_{x^N} P(x^N) \sum_{\mathcal{C} : x^N \notin J(\mathcal{C})} \Pr(\mathcal{C}). \tag{2.20}$$

The inner summation is the probability of choosing a codebook that does not well represent the x^N specified in the outer summation. The probability that a single randomly chosen codeword does not well represent x^N is

$$\Pr\left((x^N, \hat{X}^N) \notin A_\epsilon^N\right) = 1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N). \tag{2.21}$$

We emphasize here that we need to use the directed probability $\vec{P}(\hat{x}^N|x^N)$ in (2.21) because this is the distribution we used to generate the random code. Thus the probability of choosing a codebook that does not well represent x^N is

$$\left[1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N)\right]^{2^{NR}}. \tag{2.22}$$

Substituting this in (2.20), we get

$$\overline{P}_e = \sum_{x^N} P(x^N) \left[1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N)\right]^{2^{NR}}. \tag{2.23}$$

We can now use Lemma 2.7 to obtain

$$\overline{P}_e \leq \sum_{x^N} P(x^N) \left[1 - 2^{-N(I(\hat{X}\to X)+3\epsilon)} \sum_{\hat{x}^N:(x^N,\hat{x}^N)\in A_\epsilon^N} P(\hat{x}^N|x^N) \right]^{2^{NR}}. \tag{2.24}$$

As shown in [15], the inequality

$$(1-xy)^n \leq 1 - y + e^{-xn} \tag{2.25}$$

holds for $n > 0$ and $0 \leq x, y \leq 1$. Using this in (2.24), we get

$$\begin{aligned} \overline{P}_e &\leq \left[\sum_{x^N} P(x^N) \sum_{\hat{x}^N:(x^N,\hat{x}^N)\notin A_\epsilon^N} P(\hat{x}^N|x^N) \right] + e^{-2^{N(R-I(\hat{X}\to X)-3\epsilon)}} \\ &= \sum_{(x^N,\hat{x}^N)\notin A_\epsilon^N} P(x^N,\hat{x}^N) + e^{-2^{N(R-I(\hat{X}\to X)-3\epsilon)}}. \end{aligned} \tag{2.26}$$

The first term is the probability that a pair $(x^N, \hat{x}^N)$ chosen according to the distribution $P_{X^N,\hat{X}^N}$ is not directed distortion ϵ-typical. From Lemma 2.6, this vanishes as $N \to \infty$. Therefore, $\overline{P}_e \to 0$ as long as $R > I(\hat{X} \to X) + 3\epsilon$. Thus we have shown that there exists a code with rate arbitrarily close to $R^*(D)$ that has expected distortion arbitrarily close to D. □

It is worth comparing the expression in Theorem 1 for $R^*(D)$ with the optimal rate-distortion function for a source without feed-forward. The constraint set for the infimum is the same in both cases, but the objective function in $R^*(D)$ is less than or equal to that in the no-feed-forward rate-distortion function since $I(\hat{X}^N \to X^N) \leq I(\hat{X}^N; X^N)$.

We now make some observations connecting the above discussion to channel coding with feedback. Consider a channel with input X_n and output Y_n with perfect feedback, i.e. to determine X_n, the encoder knows Y^{n-1}. The channel, characterized by a sequence of distributions $\vec{P}_{\mathbf{Y}|\mathbf{X}} = \{P_{Y_n|X^n,Y^{n-1}}\}_{n=1}^{\infty}$, is fixed. What the encoder

can control is the input distribution $\vec{P}_{\mathbf{X}|\mathbf{Y}} = \{P_{X_n|X^{n-1},Y^{n-1}}\}_{n=1}^{\infty}$. Note that

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}} = \vec{P}_{\mathbf{Y}|\mathbf{X}} \cdot \vec{P}_{\mathbf{X}|\mathbf{Y}}.$$

Under the assumption that the joint process $\{X_n, Y_n\}_{n=1}^{\infty}$ is stationary and ergodic, we can use methods similar to those used in this section to show that all rates less that $sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}}} I(X \to Y)$ are achievable with feedback. Compare this with the no-feedback capacity of the channel, given by $sup_{\mathbf{P}_{\mathbf{X}}} I(X;Y)$. It is shown in [53] that when there is no feedback in the channel, $I(X;Y) = I(X \to Y)$. Hence the no-feedback capacity of the channel can be written as $sup_{\mathbf{P}_{\mathbf{X}}} I(X \to Y)$.

Comparing the expressions for capacity with and without feedback, we see that the objective function ($I(X \to Y)$) is the same; but the constraint set of optimization is larger when feedback is present since the space of $\mathbf{P}_{\mathbf{X}}$ is contained in the space of $\vec{P}_{\mathbf{X}|\mathbf{Y}}$. Compare this with the source coding problem where $\mathbf{P}_{\mathbf{X}}$ is fixed. With or without feed-forward, the constraint set of optimization remains the same ($\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ subject to distortion constraint). But the objective function with feed-forward- $I(\hat{X} \to X)$- is smaller than in the no-feed-forward case, $I(\hat{X};X)$. In summary, for channels, the boost in capacity due to feedback is due to a larger constraint set of optimization. In contrast, for sources, the decrease in the rate-distortion function due to feed-forward is due to a smaller objective function.

2.4 General sources

2.4.1 Rate-distortion theorem

In this section, we prove the rate-distortion theorem for arbitrary sources with feed-forward. We will use the method of information spectrum introduced by Han and Verdu [33]. This a powerful tool to deal with general processes without making any assumptions. Information spectrum methods have been used to derive formulas

for the capacity of general channels with and without feedback [84, 77] and the rate-distortion function of general sources [76]. They have also been used to derive error exponents for both lossless and lossy source coding of general sources [35, 39, 41, 40].

The apparatus we will use for proving coding theorems for general discrete sources with feed-forward is first described. We define a code-function, which maps the feed-forward information to a source reconstruction symbol $\hat{X}$. These code-functions are the same as the code-trees used in the previous section, but we give a formal definition here. Roughly speaking, a source code with feed-forward is a set of code-functions. The source sequence X^N determines the code-function to be used and the mapping to the reconstruction symbols is done by the decoder using the code-function and fed-forward values.

Definition 2.8. A source code-function f^N is a set of N functions $\{f_n\}_{n=1}^N$ such that $f_n : \mathcal{X}^{n-1} \to \widehat{\mathcal{X}}$ maps each source sequence $x^{n-1} \in \mathcal{X}^{n-1}$ to a reconstruction symbol $\hat{x}_n \in \widehat{\mathcal{X}}$. Denote the space of all code-functions by $\mathcal{F}^{\mathcal{N}} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \mathcal{F}_N \triangleq \{f^N : f^N$ is a code function$\}$.

Definition 2.9. A $(N, 2^{NR})$ source codebook of rate R and block length N is a set of 2^{NR} code-functions. Denote them by $f^N[w], \quad w = 1, \ldots, 2^{NR}$.
An encoder is a mapping that maps each source sequence $x^N \in \mathcal{X}^N$ to a code-function in the codebook.

Note that a source codebook and an encoder together automatically define the decoder as follows. For each source sequence of length N, the encoder sends an index to the decoder. Using the code-function corresponding to this index, the decoder maps the information fed forward from the source to produce an estimate $\hat{X}$. A code-function can be represented as a tree as in Figure 2.3. In a system

without feed forward, a code-function generates the reconstruction independent of the past source samples. In this case, the code-function reduces to a codeword. In other words, for a system without feed-forward, a source codeword is a source code-function $f^N = \{f_1, \ldots, f_N\}$ where for each $n \in \{1, \ldots, N\}$, the function f_n is a constant mapping.

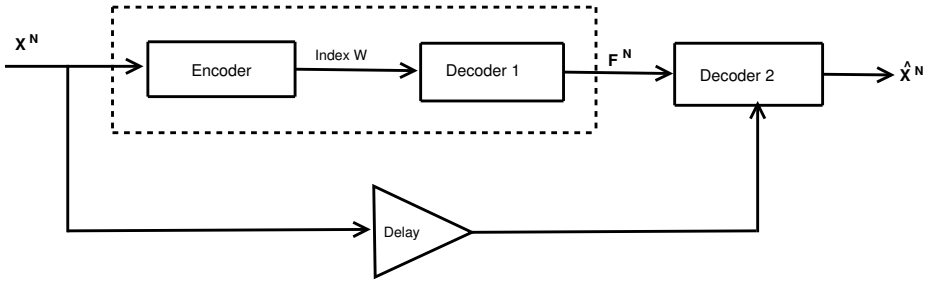

Figure 2.4: Representation of a source coding scheme with feed-forward.

A source coding system with feed-forward can be thought of as having two components. The first is the part inside the dashed box in Figure 2.4. This is a standard no-feed-forward source coding system which produces a code-function F^N as reconstruction for each source sequence X^N. In other words, for each source sequence x^N, the encoder chooses the best code-function among $f^N[W], \quad W \in \{1, \ldots, 2^{NR}\}$ and sends the index of the chosen code function. If we denote the chosen code-function by f^{*N}, the second component (decoder 2 in Figure 2.4) produces the reconstruction given by

$$\hat{X}_i = f_i^*(X^{i-1}), \qquad i = 1, \ldots, N. \tag{2.27}$$

In the sequel, we will use the notation $\hat{X}^N = f^N(X^{N-1})$ as shorthand to collectively refer to the N equations described by (2.27). In source coding with feed-forward, the encoder induces a conditional distribution $\forall f^N \in \mathcal{F}^N, x^N \in \mathcal{X}^N$ given by

$$P_{F^N|X^N}(f^N|x^N) = \begin{cases} 1, \text{ if } f^N = \text{the code-function chosen by the encoder.} \\ 0, \text{ otherwise.} \end{cases} \tag{2.28}$$

The reconstruction $\hat{x}^N$ is uniquely determined by f^N and x^N. Thus

$$P_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) = \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}. \tag{2.29}$$

Therefore, given a source distribution P_{X^N}, a source code-book and an encoder e, a unique joint distribution Q of X^N, F^N and $\hat{X}^N$ is determined: $\forall x^N \in \mathcal{X}^N, \quad f^N \in \{f^N[i] : 1 \leq i \leq 2^{NR}\}, \quad \hat{x}^N \in \hat{\mathcal{X}}^N,$

(2.30)

$$\begin{aligned} Q_{X^N,F^N,\hat{X}^N}(x^N,f^N,\hat{x}^N) &= P_{X^N}(x^N)\cdot P_{F^N|X^N}(f^N|x^N)\cdot P_{\hat{X}^N|F^N,X^N}(\hat{x}^N|f^N,x^N) \\ &= P_{X^N}(x^N)\cdot \delta_{\{f^N=e(x^N)\}}\cdot \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}, \end{aligned}$$

where $e(x^N)$ denotes the code-function chosen by the encoder for a sequence $x^N \in \mathcal{X}^N$.

We now give the general rate-distortion theorem - for arbitrary discrete sources with feed-forward without the assumptions of stationarity or ergodicity. For this, we use the machinery developed in [76] for the standard source coding problem, i.e., without feed-forward. The source distribution is a sequence of distributions denoted by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$. A conditional distribution is denoted by $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$. We consider a sequence of distortion measures $d_n(x^n,\hat{x}^n)$, and, as before, we assume $d_n(\cdot,\cdot)$ is normalized with respect to n and is uniformly bounded in n.

We give the result for two kinds of distortion criteria. The first is a constraint on the expected distortion. The second criterion is a probability of error criterion- the restriction is on the probability that the distortion is at least D. The probability of error criterion may be more useful for a general source, which may not be ergodic or stationary.

Definition 2.10 (a). *(Expected distortion criterion)* R is an ϵ-achievable rate at expected distortion D if for all sufficiently large N, there exists an $(N, 2^{NR})$ source

codebook and an associated encoder such that

$$E_{X^N}\left[d_N(x^N, \hat{x}^N)\right] \leq D + \epsilon,$$

where $\hat{x}^N$ denotes the reconstruction of x^N.

R is an achievable rate at expected distortion D if it is ϵ-achievable for every $\epsilon > 0$.

(b) *(Probability of error criterion)* R is an ϵ-achievable rate at probability-1 distortion D if for all sufficiently large N, there exists an $(N, 2^{NR})$ source codebook such that

$$P_{X^N}\left(x^N : d_N(x^N, \hat{x}^N) > D\right) < \epsilon,$$

where $\hat{x}^N$ denotes the reconstruction of x^N.

R is an achievable rate at probability-1 distortion D if it is ϵ-achievable for every $\epsilon > 0$.

We now state the definitions of a few quantities (previously defined in [84],[77]) which we will use in our coding theorems. A word about the notation used in the remainder of this paper. We will use the usual notation $P_X(x)$ to indicate the probability mass function of X evaluated at the point x. Often, we will treat the p.m.f of X as a function of the random variable X. In such situations, the function is also a random variable and we will use the notation $P(X)$ and $P_X(X)$ interchangeably to refer to this random variable.

Definition 2.11. The *limsup in probability* of a sequence of random variables $\{X_n\}$ is defined as the smallest extended real number α such that $\forall \epsilon > 0$

$$\lim_{n\to\infty} Pr[X_n \geq \alpha + \epsilon] = 0.$$

The *liminf in probability* of a sequence of random variables $\{X_n\}$ is defined as the

largest extended real number β such that $\forall \epsilon > 0$

$$\lim_{n\to\infty} Pr[X_n \leq \beta - \epsilon] = 0.$$

Definition 2.12. For any sequence of joint distributions $\{P_{X^N,\hat{X}^N}\}_{N=1}^{\infty}$, define $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \widehat{\mathcal{X}}^N$

$$i(x^N;\hat{x}^N) \triangleq \log \frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{P_{\hat{X}^N}(\hat{x}^N)P_{X^N}(x^N)} \quad , \tag{2.31}$$

$$\overline{H}(X) \triangleq \limsup_{inprob} \frac{1}{N} \log \frac{1}{P_{X^N}(X^N)} \quad , \tag{2.32}$$

$$\underline{H}(X) \triangleq \liminf_{inprob} \frac{1}{N} \log \frac{1}{P_{X^N}(X^N)} \quad , \tag{2.33}$$

$$\vec{i}(\hat{x}^N;x^N) \triangleq \log \frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)P_{X^N}(x^N)} \quad , \tag{2.34}$$

$$\overline{I}(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N}\vec{i}(\hat{X}^N;X^N) \quad , \tag{2.35}$$

$$\underline{I}(\hat{X} \to X) \triangleq \liminf_{inprob} \frac{1}{N}\vec{i}(\hat{X}^N;X^N) \quad , \tag{2.36}$$

where $\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)$ *and* $\vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)$ *are given by* (2.8) *and* (2.9) *respectively.*

We also note that the directed information from $\hat{X}^N$ to X^N can be written as

$$I(\hat{X}^N \to X^N) = \sum_{x^N,\hat{x}^N} P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)\vec{i}(\hat{x}^N;x^N). \tag{2.37}$$

As pointed out in [76], the entropy rate and the mutual information rate, defined by $\lim_{n\to\infty} \frac{1}{n} \log H(X^n)$ and $\lim_{n\to\infty} \frac{1}{n} \log I(X^n;\hat{X}^n)$ respectively, may not exist for an arbitrary random process that is neither stationary nor ergodic. But the sup-entropy rate and the inf-entropy rate ($\overline{H}(X)$ and $\underline{H}(X)$ defined above) always exist, as do the sup-information rate and the inf- information rate ($\overline{I}(X;\hat{X})$ and $\underline{I}(X;\hat{X})$ defined in [33]).

Lemma 2.13. [77] *For any sequence of joint distributions* $\{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$*, we have*

(2.38)

$$\underline{I}(\hat{X} \to X) \leq \liminf_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \leq \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \leq \overline{I}(\hat{X} \to X).$$

If

$$\underline{I}(\hat{X} \to X) = \overline{I}(\hat{X} \to X), \tag{2.39}$$

then the limit exists and all the quantities in (2.38) are equal. The class of processes for which this equality holds includes (but is not limited to) stationary and ergodic joint processes. We are now ready to state and prove the rate distortion theorem for an arbitrary source with feed-forward. In [84], Verdu and Han showed that the capacity formula for arbitrary channels without feedback is an optimization(sup) of the inf-information rate over all input distributions. Analogously, it was shown in [76] that the rate distortion function (without feed-forward) for an arbitrary source is given by an optimization(inf) of the sup-information rate. Tatikonda and Mitter [77] showed that for arbitrary channels with feedback, the capacity is an optimization of $\underline{I}(X \to Y)$, the inf-directed information rate. Our result is that the rate distortion function for an arbitrary source with feed-forward is an optimization of $\overline{I}(X \to \hat{X})$, the sup-directed information rate.

Theorem 2 (a)**.** (Expected Distortion Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with feed-forward, the infimum of all achievable rates at expected distortion D, is given by*

$$R_{ff}^{*}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}(\hat{X} \to X), \tag{2.40}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{n\to\infty} E[d_n(X^n, \hat{X}^n)]. \tag{2.41}$$

(b) (Probability of Error Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with feed-forward, the infimum of all*

*achievable rates at probability-*1 *distortion* D*, is given by*

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}(\hat{X} \to X), \tag{2.42}$$

where

(2.43)

$$\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) = \inf\left\{h : \lim_{n\to\infty} P_{X^n} P_{\hat{X}^n|X^n}\left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h\right) = 0\right\}.$$

Note that if the joint process $\{X_n, \hat{X}_n\}_{n=1}^{\infty}$ satisfies (2.39), from Lemma 2.13, the rate-distortion function becomes

$$R_{ff}(D) = \inf \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N), \tag{2.44}$$

where the infimum is evaluated according to the distortion constraint used. Although the rate-distortion function given by Theorem 2 involves optimizing a multi-letter expression involving X and $\hat{X}$, we will show in Chapter 4 that this can be evaluated in closed form for several sources and distortion measures with memory.

The detailed proofs of the direct and converse parts of Theorem 2 are found in Appendix A.2 and A.3, respectively. The proof of the direct part uses the machinery introduced in [77] for proving the capacity results for channels with feedback. The proofs for parts (a) and (b) are very similar. We only give a brief outline here of the direct coding theorem. For the sake of intuition, assume that (2.44) holds. We want to show the achievability of all rates greater than $R_{ff}(D)$ in (2.44).

Let $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^*_{\hat{X}^n|X^n}\}$ be the conditional distribution that achieves the infimum (subject to the constraint). Fix the block length N. The source code with source X^N and reconstruction F^N does not contain feed-forward (see Figure 2.4). Our goal is to construct a joint distribution over $X^N, \hat{X}^N$ and F^N, say $Q_{F^N,X^N,\hat{X}^N}$, such that the marginal over X^N and $\hat{X}^N$ satisfies

$$Q_{X^N,\hat{X}^N} = P_{X^N} P^*_{\hat{X}^N|X^N}. \tag{2.45}$$

We also impose certain additional constraints on $Q_{F^N,X^N,\hat{X}^N}$ so that [3]

$$I_Q(F^N; X^N) = I_Q(\hat{X}^N \to X^N). \tag{2.46}$$

Using (2.45) in the above equation, we get

$$I_Q(F^N; X^N) = I_{P_{X^N} P^*_{\hat{X}^N|X^N}}(\hat{X}^N \to X^N). \tag{2.47}$$

Using the usual techniques for source coding without feed-forward, it can be shown that all rates greater than $\frac{1}{N} I_Q(F^N; X^N)$ can be achieved. From (2.47), it follows that all rates greater than $I_{P_{X^N} P^*_{\hat{X}^N|X^N}}(\hat{X}^N \to X^N)$ are achievable. The bulk of the proof lies in constructing a suitable joint distribution Q.

It should be remarked here that to prove Theorem 2, we do not use the concept of directed distortion typicality introduced in Section 2.3. Notions of typicality are useful only for stationary and ergodic processes. However, when the joint process $\{X_n, \hat{X}_n\}$ is stationary and ergodic, Theorem 2(a) gives the same rate-distortion function as Theorem 1. The reason for the discussion in Section 2.3 was to give intuition about source coding with feed-forward before going into full generality.

2.4.2 Discrete memoryless sources

Consider an arbitrary discrete memoryless source (DMS). Such a source is characterized by a sequence of distributions $\{P_{X^n}\}_{n=1}^{\infty}$, where for each n, P_{X^n} is a product distribution.

We prove the following result for a DMS with expected distortion constraint and a memoryless distortion measure $d_N(x^N, \hat{x}^N) = \frac{1}{N} \sum_{i=1}^{N} d_i(x_i, \hat{x}_i)$.

Theorem 3. *Feed-forward does not decrease the rate-distortion function of a discrete memoryless source.*

---

[3] For clarity, wherever necessary, we will indicate the distribution used to calculate the information quantity as a subscript of I.

Proof. See Appendix A.4.

This result was shown in [85] for the sources that were identically distributed, in addition to being memoryless. It should be noted that Theorem 3 may not hold for a general distortion measure $d_N(x^N, \hat{x}^N)$. In other words, even when the source is memoryless, feed-forward could decrease the rate-distortion function when the distortion constraint has memory. The theorem might also not hold when the probability of error distortion constraint (Theorem 2(b)) is used instead of the expected distortion constraint regardless of the nature of the distortion measure $d_N(x^N, \hat{x}^N)$.

2.4.3 Gaussian sources with feed-forward

In this section, we study the the special case of Gaussian sources with feed-forward. A source X is Gaussian if the random process $\{X_n\}_{n=1}^{\infty}$ is jointly Gaussian. A Gaussian source is continuous valued unlike the sources hitherto discussed. However, it is straightforward to extend the results derived earlier for discrete sources to continuous sources. In particular, feed-forward does not decrease the rate-distortion function of a memoryless Gaussian source with expected mean-squared error distortion criterion. Interestingly though, feed-forward in an IID Gaussian source enables us to achieve rates arbitrarily close to the rate-distortion function with a low complexity coding scheme involving just linear processing and uniform scalar quantization (without entropy coding) at all rates [64].

An explicit characterization of the distortion-rate function for a stationary Gaussian source with feed-forward was given in [85] for an average mean-squared error distortion criterion. Here we consider arbitrary Gaussian sources and prove a result on the structure of the optimum achieving conditional distribution for any quadratic distortion criterion. As in the case of discrete memoryless sources, we use the expected distortion constraint. We now show that for a Gaussian source, $R_{ff}^*(D)$ is

achieved by a Gaussian conditional distribution.

Proposition 2.14. *Let X be an arbitrary Gaussian source with distribution $\mathbf{P_X}$. Then the optimal rate-distortion function with feed-forward with a quadratic distortion measure is achieved by a Gaussian conditional distribution.*

Proof. Suppose the conditional distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ achieves the optimal rate-distortion function. Let $\mathbf{G}_{\hat{\mathbf{X}}|\mathbf{X}} = \{G_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be a Gaussian conditional distribution such that for all N,

$$G_{X^N,\hat{X}^N} = P_{X^N} \cdot G_{\hat{X}^N|X^N}$$

is a jointly Gaussian distribution that has the same second order properties as $P_{X^N,\hat{X}^N} = P_{X^N} \cdot P_{\hat{X}^N|X^N}$. Then we will show:

1. $I_G(\hat{X}^N \to X^N) \leq I_P(\hat{X}^N \to X^N)$,
2. The average distortion is the same under both distributions, i.e.,

$$E_P[d_N(X^N, \hat{X}^N)] = E_G[d_N(X^N, \hat{X}^N)]. \tag{2.48}$$

1. We denote the densities corresponding to $P_{X^N,\hat{X}^N}$ and $G_{X^N,\hat{X}^N}$ by

$$p_{X^N,\hat{X}^N} = p_{X^N} p_{\hat{X}^N|X^N} \quad \text{and} \quad g_{X^N,\hat{X}^N} = p_{X^N} g_{\hat{X}^N|X^N}$$

Using the representation of directed information given in (2.37), we have the following

chain of inequalities

$$
\begin{aligned}
&I_P(\hat{X}^N \to X^N) - I_G(\hat{X}^N \to X^N) \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&\qquad - \int g_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{g_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&\qquad - \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{g_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)p_{X^N}(x^N)} dx^N d\hat{x}^N,
\end{aligned}
$$

where the last equality is due to the fact that $p_{X^N,\hat{X}^N}$ and $g_{X^N,\hat{X}^N}$ have the same second order properties. Continuing the chain, we have

$$
\begin{aligned}
&I_P(\hat{X}^N \to X^N) - I_G(\hat{X}^N \to X^N) \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N,\hat{x}^N)\vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{g_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{\vec{p}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)}{\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{\vec{p}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N,\hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{p'_{X^N,\hat{X}^N}(x^N,\hat{x}^N)} dx^N d\hat{x}^N,
\end{aligned}
$$

where $p'_{X^N,\hat{X}^N}$ is the joint distribution $\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \cdot \vec{p}_{\hat{X}^N|X^N}$. Then last expression is the Kullback-Leibler distance between the distributions p and p' and is thus non-negative.

2. Since $P_{X^N,\hat{X}^N}$ and $G_{X^N,\hat{X}^N}$ have the same second order properties, it follows that the expected distortion is the same under both distributions. □

Thus for Gaussian sources with a quadratic distortion measure, the optimizing conditional distribution can be taken to be jointly Gaussian. We also have the following result from [77] for jointly Gaussian distributions. For any jointly Gaussian

distribution $\mathbf{P}_{\mathbf{X^N},\hat{\mathbf{X}}^\mathbf{N}} = \{P_{X^N,\hat{X}^N}\}_{n=1}^{\infty}$,

$$\overline{I}(\hat{X} \to X) = \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N). \tag{2.49}$$

This property follows from the asymptotic equipartition property, which is valid for an arbitrary Gaussian random processes (Theorem 5, [18]). Thus the rate-distortion function for an arbitrary Gaussian source with expected mean-squared error distortion criterion can be written as

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\le D} \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N), \tag{2.50}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) = \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2] \tag{2.51}$$

and $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ can be taken to be Gaussian.

2.5 Error exponents

In this section we study the error exponent for source coding with feed-forward. The error-exponent for lossy, fixed-length source coding of a stationary memoryless source (without feed-forward) with a single-letter distortion measure was derived by Marton [52] and by Blahut[8]. Recently, Iriyama derived the error exponent for lossy, fixed-length coding of a general source without feed-forward with a general distortion measure [40]. For lossless source coding, the reliability function for fixed-length coding of a general source was first studied in [35] and then in [41, 39]. Error exponents for lossy/lossless coding of certain classes of discrete sources were earlier studied in [57, 49, 80, 55, 3].

The error exponent for fixed-length lossy source coding with feed-forward is derived in [85] for sources that can be auto-regressively represented with an IID innovations process and is shown to be the same as Marton's no-feed-forward error

exponent [52]. In this section, we will use the approach and framework of [40] to obtain a formula for the error exponent for fixed-length lossy coding of any general source with feed-forward.

For a source-code with feed-forward of block-length N, let $\epsilon_N(D)$ be the probability of the distortion exceeding D.

$$\epsilon_N(D) = \Pr\left\{d_N(X^N, \hat{X}^N) > D\right\}. \tag{2.52}$$

We want to determine the infimum of all achievable coding rates such that asymptotically $\epsilon_N(D) \sim e^{-Nr} \quad (N \to \infty)$. This is called the *minimum (D, r)-achievable rate with feed-forward for the source* $\mathbf{X}$ *and is denoted* $R_{ff}(D, r|\mathbf{X})$. We will derive a formula for this in the following section.

2.5.1 'Good' source codes with feed-forward

In this section, we will determine the minimum (D, r)-achievable rate for the source $\mathbf{X}$ with feed-forward. Defining the problem formally, consider a sequence of $\left\{(N, 2^{NR_N})\right\}_{N=1}^{\infty}$ source codes with feed-forward. Each code in this sequence is defined according to Definition 2.1. We are interested in a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward such that

$$\begin{aligned} \limsup_{N\to\infty} R_N &\leq R \quad \text{and} \\ \liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{\epsilon_N(D)} &\geq r. \end{aligned} \tag{2.53}$$

Definition 2.15. *[40]* The minimum (D, r)-achievable rate for the source $\mathbf{X}$ with feed-forward is defined as

$$R_{ff}(D, r|\mathbf{X}) = \inf\left\{R : \exists \quad \textit{a sequence of codes satisfying } (2.53)\right\}. \tag{2.54}$$

The minimum (D, r)-achievable rate will be expressed in terms of a rate-distortion function with feed-forward. This rate-distortion function is defined according to a

distortion constraint that is different from those considered in Theorem 2. This is described as follows. Consider a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward satisfying

$$\begin{aligned} &\limsup_{N\to\infty} R_N \leq R \quad \text{and} \\ &\limsup_{N\to\infty} (1 - \epsilon_N(D)) > 0. \end{aligned} \tag{2.55}$$

In other words, we are interested in a sequence of codes with rate R. Further, the probability of correct decoding should be non-zero for infinitely many codes in this sequence. We will need the rate-distortion function with feed-forward defined according to this criterion.

Definition 2.16. *[40]* The Rate-distortion function $R^*_{ff}(D|\mathbf{X})$ for the source $\mathbf{X}$ with feed-forward is defined as

$$R^*_{ff}(D|\mathbf{X}) = \inf\{R : \exists \text{ a sequence of codes satisfying } (2.55)\}. \tag{2.56}$$

Finally, we will need a couple of divergence quantities to express the minimum (D, r)-achievable rate. We have $D_u(\mathbf{Y}||\mathbf{X})$ and $D_l(\mathbf{Y}||\mathbf{X})$ defined by

$$\begin{aligned} D_u(\mathbf{Y}||\mathbf{X}) &= \limsup_{n\to\infty} \frac{1}{n} D(Y^n||X^n) \\ D_l(\mathbf{Y}||\mathbf{X}) &= \liminf_{n\to\infty} \frac{1}{n} D(Y^n||X^n) \end{aligned} \tag{2.57}$$

We can now state our result.

Theorem 4. *For any $D, r > 0$,*

$$\sup_{\mathbf{Y}: D_l(\mathbf{Y}||\mathbf{X})<r} R^*_{ff}(D|\mathbf{Y}) \leq R_{ff}(D, r|\mathbf{X}) \leq \sup_{\mathbf{Y}: D_l(\mathbf{Y}||\mathbf{X})\leq r} R^*_{ff}(D|\mathbf{Y}), \tag{2.58}$$

with equalities if $R_{ff}(D, r|\mathbf{X})$ is continuous at r. Further,

(2.59)

$$\inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D} \underline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}) \leq R^*_{ff}(D|\mathbf{Y}) \leq \inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D_1} \underline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}), \quad 0 < D_1 < D$$

with equalities if continuous at D.

Proof. In Appendix A.5.

Let us examine the case when $R_{ff}(D, r|\mathbf{X})$ is continuous. Then the minimum (D, r)-achievable rate can be expressed as

$$R_{ff}(D, r|\mathbf{X}) = \sup_{\mathbf{Y}: D_l(\mathbf{Y}||\mathbf{X}) \leq r} R^*_{ff}(D|\mathbf{Y}). \tag{2.60}$$

This can be pictured in a manner analogous to the interpretation of the error exponent for stationary memoryless sources using the type-covering lemma [5, 19]. Loosely speaking, for the error decay with exponent r, we need the code to cover all sequences belonging to source distributions that are at a distance within r from the 'true' distribution $\mathbf{P_X}$. This is possible if we build a code with rate given by (2.60).

We observe that the minimum (D, r) achievable rate increases with r. As we should expect, we also see that it approaches the feed-forward rate-distortion function of $\mathbf{X}$ as r approaches 0.

From (2.60), it is also clear that the minimum (D, r)-achievable rate for a source with feed-forward is smaller than for the same source without feed-forward. Without feed-forward, the formula is the supremum of the no-feed-forward rate-distortion function $R^*(D|\mathbf{Y})$ which is clearly greater than the corresponding fed-forward rate-distortion function $R^*_{ff}(D|\mathbf{Y})$.

2.5.2 'Bad' source codes with feed-forward

If the coding rate is sufficiently small, then the probability $\epsilon_N(D)$ tends to one. Similar to [40], we can study the performance of bad feed-forward codes. In this section, we will determine the minimum coding rate $R^*_{ff}(D, r|\mathbf{X})$ for which the probability of distortion being less than or equal to D goes to zero exponentially fast with exponent r. We are interested in a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward

such that

$$\begin{aligned} &\limsup_{N\to\infty} R_N \le R \quad \text{and} \\ &\liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{1-\epsilon_N(D)} \le r. \end{aligned} \tag{2.61}$$

We define a minimum achievable rate with feed-forward $R^*_{ff}(D, r|\mathbf{X})$

$$R^*_{ff}(D, r|\mathbf{X}) = \inf \{R : \exists \text{ a sequence of codes satisfying } (2.61)\}. \tag{2.62}$$

We will express $R^*_{ff}(D, r|\mathbf{X})$ in terms of the rate-distortion function defined as follows. Consider a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward satisfying

$$\begin{aligned} &\limsup_{N\to\infty} R_N \le R \quad \text{and} \\ &\limsup_{N\to\infty} \epsilon_N(D) = 0. \end{aligned} \tag{2.63}$$

This condition is similar to (but not the same as) the probability-1 distortion constraint. For a source $\mathbf{Y}$, we define

$$R_{ff}(D|\mathbf{Y}) = \inf \{R : \exists \text{ a sequence of codes satisfying } (2.63)\}. \tag{2.64}$$

We are now ready to state our result in terms of $R_{ff}(D|\mathbf{Y})$.

Theorem 5. *For any $D, r > 0$,*

$$R^*_{ff}(D, r|\mathbf{X}) = \inf_{\mathbf{Y}: D_u(\mathbf{Y}||\mathbf{X}) \le r} R_{ff}(D|\mathbf{Y}). \tag{2.65}$$

Further,

(2.66)

$$\inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}}) \le D} \overline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}) \le R_{ff}(D|\mathbf{Y}) \le \inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}}) \le D_1} \overline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}), \quad 0 < D_1 < D$$

with equalities if continuous at D.

The proof of this theorem is found in Appendix A.5 along with the proof of Theorem 4.

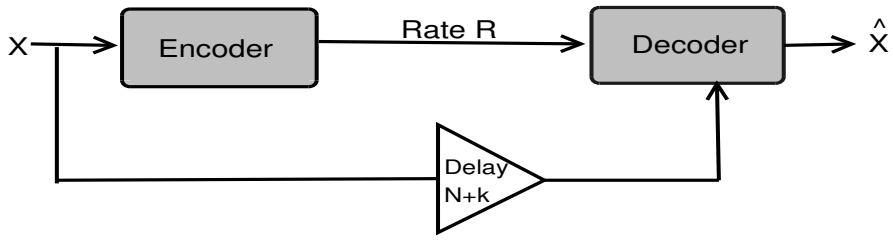

Figure 2.5: Source coding system with k-delayed feed-forward.

2.6 Feed-Forward with arbitrary delay

Recall from the discussion in Chapter 1 that our problem of source coding with noiseless feed-forward is meaningful for any delay larger than the block length N. Our results in the preceding sections assumed that the delay was $N+1$, i.e., to reconstruct the ith sample the decoder had perfect knowledge of first $n-1$ samples.

We now extend our results for a general delay $N+k$, where N is the block length. We call this delay k feed-forward. Figure 2.5 shows a system with delay k feed-forward. The encoder is a mapping to an index set: $e : \mathcal{X}^N \to \{1, \ldots, 2^{NR}\}$. The decoder receives the index transmitted by the encoder, and to reconstruct the nth sample, it has access to all the past $(n-k)$ samples of the source. In other words, the decoder is a sequence of mappings $g_n : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{n-k} \to \widehat{\mathcal{X}}, \quad n = 1, \ldots, N$.

The key to understanding feed-forward with arbitrary delay is the interpretation of directed information in Section 2.2.2. Recall from (2.3) that the directed information can be expressed as

$$I(\hat{X}^N \to X^N) = I(\hat{X}^N; X^N) - \sum_{n=2}^{N} I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}). \tag{2.67}$$

With delay k feed-forward, the decoder knows X^{i-k} to reconstruct $\hat{X}_n$. Here, we need not spend $I(X^{n-k}; \hat{X}_n | \hat{X}^{n-1})$ bits to code this information, hence this rate comes for free. In other words, the performance limit on this problem is given by the minimum

of

$$I_k(\hat{X}^N \to X^N) \triangleq I(\hat{X}^N; X^N) - \sum_{n=k+1}^{N} I(X^{n-k}; \hat{X}_i | \hat{X}^{n-1}) \tag{2.68}$$

$$\stackrel{(a)}{=} \sum_{n=1}^{N} I(\hat{X}^{n+k-1}; X_n | X^{n-1}), \tag{2.69}$$

where (a) is proved in Appendix A.6.

Observing (2.68), we make the following comment. In any source coding problem, the mutual information $I(\hat{X}^N; X^N)$ is the fundamental quantity to characterize the rate-distortion function. With feed-forward, we get some information for free and the rate-distortion function is reduced by a quantity equal to the 'free information'. One can use very similar arguments to characterize the capacity of channels with feedback delay $k \geq 1$.

We now state the rate-distortion theorem for feed-forward with general delay. We omit the proof since it is similar to the ones in the preceding sections.

Definition 2.17.

$$\vec{P}^k(\hat{X}^N | X^N) \triangleq \prod_{n=1}^{N} P(\hat{X}_n | \hat{X}^{n-1}, X^{n-k}), \tag{2.70}$$

$$I_k(\hat{X}^N \to X^N) \triangleq \sum_{n=1}^{N} I(\hat{X}^{n+k-1}; X_n | X^{n-1}) \tag{2.71}$$

$$= \sum_{x^N, \hat{x}^N} P_{X^N, \hat{X}^N}(x^N, \hat{x}^N) \log \frac{P_{X^N, \hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N) \vec{P}^k(\hat{X}^N | X^N)},$$

$$\overline{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N} \log \frac{P_{X^N, \hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N) \vec{P}_k(\hat{X}^N | X^N)}. \tag{2.72}$$

Theorem 6 (Rate-Distortion Theorem)**.**

(a) (Expected Distortion Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with delay k feed-forward, the infimum*

of all achievable rates at expected distortion D, is given by

$$R_{ff}^*(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}_k(\hat{X} \to X), \tag{2.73}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{n\to\infty} E[d_n(X^n, \hat{X}^n)]. \tag{2.74}$$

(b) (Probability of Error Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P}_{\mathbf{X}}$, the rate-distortion function with delay k feed-forward, the infimum of all achievable rates at probability-*1 *distortion D, is given by*

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}_k(\hat{X} \to X), \tag{2.75}$$

where

(2.76)

$$\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) = \inf\left\{h : \lim_{n\to\infty} P_{X^n,\hat{X}^n}\left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h\right) = 0\right\}.$$

We can also extend the error exponent results of Theorems 4 and 5 to feed-forward with arbitrary delay. As a final remark, we note that when $k \to \infty$, the problem becomes source coding without feed-forward. As we would expect, the delay k feed-forward rate-distortion function given by Theorem 6 reduces to the no-feed-forward rate distortion function $\inf \bar{I}(\hat{X}; X)$ as $k \to \infty$.

2.7 Conclusion

In this chapter, we defined and analyzed a source coding model with feed-forward. This is a source coding system in which the decoder has knowledge of previous source samples while reconstructing the present sample. This problem was first considered in [85] where the distortion-rate function was characterized for a class of sources.

We have derived the optimal rate-distortion function for a general source with feed-forward. We also characterized the error exponent for a general source with feed-forward. Specifically, for a source to be encoded with distortion D, we found the minimum rate at which the probability of error decays with exponent r.

We presented an intuitive interpretation of the role of directed information in source coding with feed-forward. Guided by this interpretation, we generalized the definition of directed information in order to analyze the feed-forward model with an arbitrary delay. The problem of source coding with feed-forward can be considered the dual of channel coding with feedback. In Chapter 4, we demonstrate that the rate-distortion function with feed-forward can be evaluated in closed-form for several sources and distortion measures with memory.

In this chapter, we considered fixed length block coding with feed-forward. An interesting extension would be to consider feed-forward with variable-rate lossy source coding. With variable rates, using techniques similar to Han [34], it should be possible to code at rates lower than the $\limsup_{in\ prob}$ of the directed information.

CHAPTER 3

Directed Information for Channel Coding with Delayed Feedback and Common State Information

3.1 Introduction

Feedback is widely used in communication systems to help combat the effect of channel noise. It is a well-known result in information theory that feedback does not increase the capacity of a discrete memoryless channel [74]. For point-to-point channels with memory, feedback can indeed increase the capacity. Marko was one of the first to consider the tools required to address this problem [50] by introducing a directed notion of information quantities. Inspired by Marko's work, Massey [53] introduced the concept of directed information flowing from a random sequence X^N to a random sequence Y^N, and showed that it can be used to upper bound the feedback capacity. Tatikonda later established the capacity of a general single-user channel in terms of the directed information flowing from the input to the output [77]. All these works considered feedback with delay 1 from the receiver to the transmitter. In other words, to generate the channel input at time n, the encoder knows the channel outputs until time $n-1$. Tatikonda and Mitter [78] also established that the feedback capacity of a channel with $k-$delayed feedback could be characterized using essentially the same framework as for feedback delay 1. The problem of characterizing

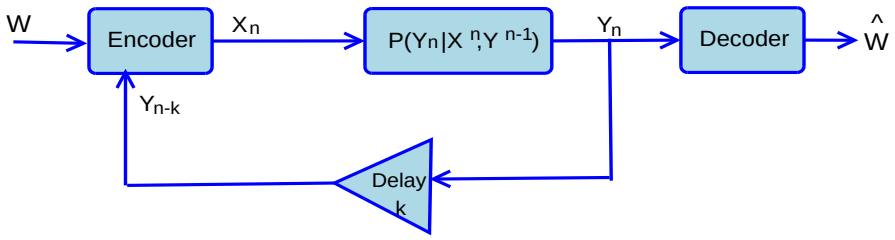

Figure 3.1: Channel with k-delayed feedback

the feedback capacity of channels has been studied by several authors in a variety of settings. An incomplete list includes [18, 53, 46, 77, 44, 45]. We must also mention that channel coding with feedback is closely related to source coding with feed-forward and can be considered its dual.

After a brief review of some results on feedback capacity, we present an interpretation of directed information in Section 3.2. This yields some insight about its relevance in the context of feedback capacity. We use this intuition in Section 3.3 to characterize the capacity of channels with delayed feedback and channel side-information which is available with some delay at both encoder and decoder. In Section 3.4, we consider a system where a source has to be transmitted over a channel with delay l feedback. The source is reconstructed with delay k feed-forward to a specified distortion level. Under suitable conditions, we show that source-channel separation holds.

3.2 Channel Capacity with delayed feedback

Consider a channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Let X_n and Y_n denote the channel input and output at time n, respectively. The channel is assumed to have noiseless feedback with delay k ($k \geq 1$), as shown in Figure 3.1. This means that at time instant n, the encoder has perfect knowledge of the channel outputs until time $n-k$ to produce the channel input X_n. We first define the ingredients

needed to specify a channel coding problem with feedback delay k.

Definition 3.1. *(a)* A channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined as a sequence of conditional distributions $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}} \triangleq \{P^{ch}_{Y_n|X^n,Y^{n-1}}\}_{n=1}^{\infty}$.
(b) An $(N, 2^{NR})$ channel code (block length N, rate R) for a channel with feedback delay k consists of a sequence of encoder mappings $e_i, i = 1, \ldots, N$ and a decoder g, where

$$e_i : \{1, \ldots, 2^{NR}\} \times \mathcal{Y}^{i-k} \to \mathcal{X}, \quad i = 1, \ldots, N$$

$$g : \mathcal{Y}^N \to \{1, \ldots, 2^{NR}\}.$$

If $W \in \{1, \ldots, 2^{NR}\}$ denotes the transmitted message, the channel input at time i is given by $X_i = e_i(W, Y^{i-k})$ for $i > k$, and $e_i(W)$ for $i \leq k$. The decoder reconstructs the message as $\hat{W} = g(Y^N)$.

The probability of error and achievable rates are defined in the usual way. If W is the message (with uniform distribution over the set $\{1, 2, \ldots, 2^{NR}\}$) that was transmitted, then

$$P_e = Pr(g(Y^N) \neq W).$$

Definition 3.2. R is an ϵ-achievable rate with $k-$delay feedback if for all sufficiently large N, there exists an $(N, 2^{NR})$ channel code such that $P_e < \epsilon$.
A rate R is achievable if it is ϵ-achievable for every $\epsilon > 0$. The infimum of all achievable rates is the $k-$delay feedback capacity C^k_{fb}.

Consider an input distribution for a channel (with distribution $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$) with $k-$delay feedback as a sequence of distributions of the form

$$\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}} = \{P_{X_n|X^{n-1},Y^{n-k}}\}_{n=1}^{\infty}. \tag{3.1}$$

Given an equiprobable distribution on the set of messages, any channel code (as defined in Definition 3.1) will correspond to a unique input distribution. For

a channel with $k-$delay feedback, consider a time-line of how input symbols are produced at the encoder until time N.

$$X_1 \quad X_2 \quad \dots \quad X_k \quad X_{k+1}(Y^1) \quad X_{k+1}(Y^2) \quad \dots \quad X_N(Y^{N-k}).$$

The input distribution of the system until time N is given by

$$\left\{P_{X_1}, P_{X_2|X_1}, \dots, P_{X_k|X^{k-1}}, P_{X_k|X^{k-1},Y^1}, \dots, P_{X_N|X^{N-1},Y^{N-k}}\right\}.$$

This, coupled with the channel distribution $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$, specifies the joint distribution of the input and output at time N as

(3.2)

$$\begin{aligned} P_{X^N,Y^N} &= P_{X_1} \cdot P^{ch}_{Y_1|X_1} \dots P_{X_{k+1}|X^k,Y^1} \cdot P^{ch}_{Y_{k+1}|X^{k+1},Y^k} \dots P_{X_N|X^{N-1},Y^{N-k}} \cdot P^{ch}_{Y_N|X^N,Y^{N-k}} \\ &= \vec{P}^{k}_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}, \end{aligned}$$

where

$$\vec{P}^{k}_{X^N|Y^N} = \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k}}, \tag{3.3}$$

$$\vec{P}^{ch}_{Y^N|X^N} = \prod_{n=1}^{N} P^{ch}_{Y_n|X^n,Y^{n-1}}. \tag{3.4}$$

Since no assumptions (such as stationarity) are made on the joint distribution of the input and output, the results in [77, 78] require the use of directed information spectrum, defined as follows.

For any sequence $\{P_{X^N,Y^N}\}_{N=1}^{\infty}$ of joint distributions on the input and output (with P_{X^N,Y^N} as in (3.2)), define $\forall x^N \in \mathcal{X}^N, y^N \in \mathcal{Y}^N$,

$$\vec{i}(x^N;y^N) \triangleq \log \frac{P_{X^N,Y^N}(x^N,y^N)}{\vec{P}^{1}_{X^N|Y^N}(x^N|y^N)P_{Y^N}(y^N)} \quad , \tag{3.5}$$

$$\underline{I}(X \to Y) \triangleq \liminf_{inprob} \frac{1}{N}\vec{i}(X^N;Y^N), \tag{3.6}$$

where $\vec{P}^{1}_{X^N|Y^N}P_{Y^N}$ is defined by (3.3). We now restate the result from [78].

Theorem 7. [78] *For a general channel* $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ *, the capacity with* $k-$*delay feedback is given by*

$$C^k_{fb} = \sup_{\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}} \underline{I}(X \to Y), \tag{3.7}$$

For comparison, we also reproduce the general formula for the no-feedback capacity given by Verdu and Han [84]. For a channel without feedback, the set of allowable input distributions is

$$\mathbf{P_X} = \{P_{X_n|X^{n-1}}\}_{n=1}^{\infty}.$$

Theorem 8. [84] *For a general channel* $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ *, the no-feedback capacity is given by*

$$C_{no-fb} = \sup_{\mathbf{P_X}} \underline{I}(X;Y), \tag{3.8}$$

3.2.1 Intuition

In this subsection, we give an interpretation of directed information in the context of channel coding with feedback. For this subsection alone, we will assume that the joint input-output process $\{X^n, Y^n\}_{n=1}^{\infty}$ is stationary and ergodic. This is only to keep the expressions intuitive and to give insight into the feedback problem.

We start with a channel without feedback. Under the stationary, ergodic assumption, the no-feedback capacity[1] from Theorem (8) is [2]

$$C_{no-fb} = \sup_{\mathbf{P_X}} \lim_{N\to\infty} \frac{1}{N} I(X^N; Y^N). \tag{3.9}$$

When there is no feedback, the interpretation is that $I(X^N; Y^N)$ is the reduction in uncertainty of the input X^N when the decoder observes Y^N. When there is feedback

[1] In this subsection alone, the term 'capacity' is used loosely, i.e., to denote the maximum achievable rate assuming the joint process $\mathbf{X}, \mathbf{Y}$ is stationary and ergodic.

[2] also [61, 79]

with delay k, to generate the input X_n, the encoder knows all the past outputs Y^{n-k}. Hence the information $I(Y^{n-k}; X_n|X^{n-1})$ is already known at both encoder and decoder due to the feedback and is not 'actually transmitted'. In light of this interpretation, the mutual information $I(X^N; Y^N)$ is still the fundamental quantity that characterizes the capacity, but the information that is known at both ends due to the feedback $\left(\sum_{n=k+1}^{N} I(Y^{n-k}; X_n|X^{n-1})\right)$ should be subtracted out. Thus we expect that the capacity with delay k feedback is characterized by

$$\begin{aligned} I_k(X^N \to Y^N) &= \left[I(X^N; Y^N) - \sum_{n=k+1}^{N} I(Y^{n-k}; X_n|X^{n-1}) \right], \\ &= \sum_{n=1}^{N} I(X^{n+k-1}; Y_n|Y^{n-1}) \end{aligned} \tag{3.10}$$

optimized over the space of input distributions $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$. In (3.10), the second equality follows from (2.69).

We will now show that for channels with feedback delay k, the nature of the input distribution ensures that the k-directed information is equal to the directed information (as defined in (2.3)). This is true for all $k \geq 1$. We will first need the following lemma, which gives another representation of k-directed information.

Lemma 3.3.

$$I_k(X^N \to Y^N) = \sum_{x^N, y^N} P(x^N, y^N) \log \frac{P(x^N, y^N)}{\vec{P}^k(x^N|y^N) \cdot P(y^N)}.$$

Proof. In Appendix B.1.

Proposition 3.4. *For any channel with delay k feedback ($k \geq 1$),*

$$I_k(X^N \to Y^N) = I(X^N \to Y^N), \quad \forall N. \tag{3.11}$$

Proof. As explained in the previous subsection (see (3.2),(3.3) and(3.4)), the joint distribution of the input and the output until time N in a channel with delay k

feedback is

$$P_{X^N,Y^N} = \vec{P}^{k}_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}, \tag{3.12}$$

We also observe that, using Bayes' rule, the joint probability distribution of the input and output at time N can always be written as

$$P_{X^N,Y^N} = \prod_{n=1}^{N} P_{X_n,Y_n|X^{n-1},Y^{n-1}} = \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1}} \cdot P_{Y_n|X^n,Y^{n-1}} = \vec{P}^{1}_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}. \tag{3.13}$$

From (3.12) and (3.13), we conclude that for a channel with delay k feedback,

$$\vec{P}^{k}_{X^N|Y^N} = \vec{P}^{1}_{X^N|Y^N}. \tag{3.14}$$

Using this with Lemma 3.3, we obtain

$$I_k(X^N \to Y^N) = I(X^N \to Y^N)$$

for any channel with delay k feedback. □

Using the arguments leading to (3.10) and Proposition 3.4, we have intuitively justified that the k-delay feedback capacity should be characterized by the limit of the directed information $I(X^N \to Y^N)$, optimized over the space of all input distributions $\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}}$. This explains the capacity formula in Theorem 7, which is valid for arbitrary channels and input distributions.

In summary, for any feedback delay k, the constraint on the input distribution ensures that

$$\vec{P}^{k}_{X^N|Y^N} = \vec{P}^{1}_{X^N|Y^N}.$$

holds.[3] This makes the objective function equal to the directed information for all k. But the constraint set of optimization ($\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}}$) gets progressively smaller as k

[3]The no feedback case is a special case of delayed feedback (with delay $k = \infty$). When there is no feedback, the input distribution satisfies $P_{X^N} = \vec{P}^{1}_{X^N|Y^N}$.

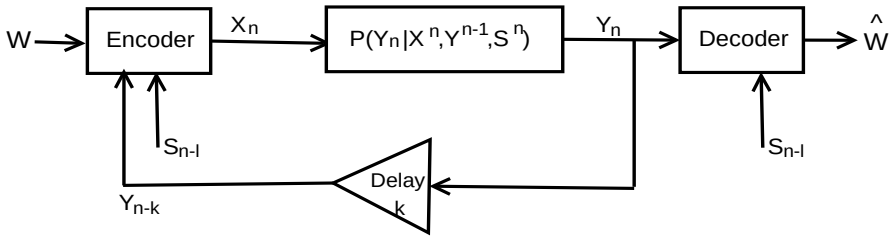

Figure 3.2: Channel with k-delayed feedback and l-delayed side-information

increases, i.e.,

$$\left\{\mathbf{P}^1_{\mathbf{X}|\mathbf{Y}}\right\} \supset \left\{\mathbf{P}^2_{\mathbf{X}|\mathbf{Y}}\right\} \supset \ldots \left\{\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}\right\} \supset \ldots \supset \{\mathbf{P}_{\mathbf{X}}\}$$

from unit-delay feedback ($\mathbf{P}^1_{\mathbf{X}|\mathbf{Y}}$) to k-delay feedback ($\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$) to no feedback ($P_{\mathbf{X}}$). Thus for channels, the boost in capacity of channels due to feedback is because of a larger constraint set available to optimize the same objective function.

In contrast, for sources with feed-forward, a result similar to Proposition 3.4 is not true. For source coding with feed-forward, observe from Theorem 6 and the discussion in Section 2.6 that for all delays k, the constraint set is the same- the set of all conditional distributions satisfying the distortion constraint. The decrease in the rate-distortion function due to feed-forward is because we optimize a smaller objective function; the objective function $I_k(\hat{X} \to X)$ is smallest for $k = 1$ and gets progressively larger with increasing k.

3.3 Channels with feedback and side-information

We now use the interpretation presented in the previous section to study a channel coding problem with causal/non-causal side-information in addition to feedback. This side-information is available at both the encoder and the decoder with some delay l, which could be different from the feedback delay k. To obtain the capacity expression for this problem, we will use the interpretation of directed information in

the context of channel coding with feedback presented in the previous section. These can then be formally proved.

We analyze the model shown in Figure 3.2. There is a channel with input X, channel side S and output Y. It is defined by the sequence of distributions

$$\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X},\mathbf{S}} = \{P_{Y_n|X^n,Y^{n-1},S^n}\}_{n=1}^{\infty}. \tag{3.15}$$

The side information is characterized by $\mathbf{P_S} = \{P_{S^n}\}$ and is assumed to be produced independent of the channel, encoder and decoder. The variable S_n is a random variable that can affect the behavior of the channel from time n onwards. Hence it is reasonable to stipulate that at any time n, given the inputs until time n and all the past channel outputs, the output Y_n depends only on the side information samples until time n, i.e.,

$$P_{Y_n|X^n,Y^{n-1},S^N} = P_{Y_n|X^n,Y^{n-1},S^n}, \quad \forall n. \tag{3.16}$$

There is k-delayed feedback, as in the previous section. In addition, the side information is known with delay l at *both* the encoder and decoder. Thus at time n, the encoder has Y^{n-k} and S^{n-l} available to generate channel input X_n.

An $(N, 2^{NR})$ channel code (block length N, rate R) for a channel with feedback delay k consists of a sequence of encoder mappings $e_i, i = 1, \ldots, N$ and a decoder g, where

$$\begin{aligned} e_i &: \{1, \ldots, 2^{NR}\} \times \mathcal{Y}^{i-k} \times \mathcal{S}^{i-l} \to \mathcal{X}, \quad i = 1, \ldots, N \\ g &: \mathcal{Y}^N \times \mathcal{S}^N \to \{1, \ldots, 2^{NR}\}. \end{aligned} \tag{3.17}$$

The probability of error, achievable rates and capacity are defined in the natural way. We allow the possibility that l could be negative, i.e., the channel side information is available non-causally. For instance, $l = -3$ means that at time n, side symbols

S^{n+3} are available to both encoder and decoder. It is understood that for negative l, $S^{n-l} = S^N$ when $n - l \geq N$. Since the channel input can depend on the fed-back symbols and the available side information, the input distribution is of the form

$$\mathbf{P}^{k,l}_{\mathbf{X}|\mathbf{Y},\mathbf{S}} = \{P_{X_n|X^{n-1},Y^{n-k},S^{n-l}}\}_{n=1}^{\infty}. \tag{3.18}$$

Causal availability of side-information: This is the case with $l \geq 0$. Consider N uses of the channel. First, we assume that the variables S^N are produced *a priori* randomly according to $\{P_{S_n|S^{n-1}}\}_{n=1}^{\infty}$. Then, the joint distribution at time N is given by

$$\begin{aligned} P_{X^N,Y^N,S^N} &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^N} \cdot P_{Y_n|Y^{n-1},X^n,S^N} \\ &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{3.19}$$

where we have made two practical assumptions to obtain the second equality. The first one is the physical constraint on the channel input distribution that at time n, it can produce input X_n based only on what is available to it, viz., $(X^{n-1}, Y^{n-k}, S^{n-l})$. The second assumption stems from the definition of the channel.

On the other hand, with causal side information, we can also assume that the S_n are produced in real time, i.e., at time n, S_n, X_n are produced and the channel acts on S^n and X^n to produce Y_n. Then, the joint distribution at time N is determined as

$$\begin{aligned} P_{X^N,Y^N,S^N} &= \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-1},S^n} \cdot P_{Y_n|Y^{n-1},X^n,S^n} \\ &= \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{3.20}$$

where we have again made the assumption relating to the physical constraints on the channel input. (3.19) and (3.20) arise from two different physical models that

result in the same joint distribution because of the practical assumptions inherent in each case. Without loss of generality, we will consider the first model where S^N is produced *a priori*, since this model can also be used when the side information is known non-causally.

Non-causal availability of side information: Here, the delay l is negative. If we denote $m = -l$, the channel side information is produced *a priori* and the side information available at the encoder and decoder at time n is S^{n+m}, with $m > 0$. Hence, to generate input X_n, the encoder can use m samples of S from the future. The joint distribution at time N is given by

$$\begin{aligned} P_{X^N,Y^N,S^N} &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^N} \cdot P_{Y_n|Y^{n-1},X^n,S^N} \\ &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^{n+m}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{3.21}$$

where we have used the physical constraint on the channel input distribution and the definition of the channel to obtain the second equality. Noting that $m = -l$, we see from (3.19), (3.20) and (3.21) that for all side information delays l, the joint distribution until time N is

$$P_{X^N,Y^N,S^N} = \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}.$$

3.3.1 Intuition

For all values of l, we can think of the side information symbols $\{S_n\}$ as additional outputs of the channel that are 'fed back' to the encoder with delay l. This is justified since the side information is available with delay l at the encoder and the decoder acts on the channel outputs and side information only at the end of all reception (at time $N + l$). Recall that the encoder knows Y^{n-k} and S^{n-l} to produce the input X_n at time n and the decoder reconstructs the message using Y^N, S^N.

Suppose now that the decoder received the side information, but the encoder had access to neither the side information nor the channel feedback. Then the capacity[4] would be characterized by $I(X^N; Y^N S^N)$ since Y^N, S^N can be now considered channel outputs and there is no feedback to the encoder. The interpretation is that $I(X^N; Y^N S^N)$ is the reduction in uncertainty of the input X^N when the decoder observes both Y^N and S^N.

In our problem, there is feedback of both the channel output and the side-information, with delays k and l, respectively. Since the past outputs Y^{n-k} and the side-information samples S^{n-l} are already known to the encoder to generate X_n, the information $I(Y^{n-k}S^{n-l}; X_n|X^{n-1})$ is not 'really transmitted' and comes for free. So the capacity of a channel whose output Y is fed back with delay k and side-information S is available with (possibly negative) delay l at both the transmitter and the receiver must be characterized by

$$I_{k,l}(X^N \to Y^N S^N) = I(X^N; Y^N S^N) - \sum_{n=\min(k,l)+1} I(X_n; Y^{n-k}, S^{n-l}|X^{n-1}) \tag{3.22}$$

optimized over the space of input distributions $\mathbf{P}^{k,l}_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$. In the same spirit as subsection 3.2.1, we will now see that the nature of the input distribution ensures that the objective function $I_{k,l}(\hat{X}^N \to X^N S^N)$ is the same for all delays l and k. We will first need the following representation of $I_{k,l}(X^N \to Y^N S^N)$.

Lemma 3.5.

$$I_{k,l}(X^N \to Y^N S^N) = E\left[\frac{1}{n}\log\frac{P_{X^N,Y^N,S^N}}{\vec{P}^{k,l}_{X^N|Y^N,S^N}\cdot P_{Y^N,S^N}}\right], \tag{3.23}$$

[4]In this subsection alone, we assume that the joint process $(\mathbf{X}, \mathbf{Y}, \mathbf{S})$ is jointly stationary and ergodic. The term 'capacity' is used loosely- to denote the maximum achievable rate under this assumption.

where

$$\vec{P}^{k,l}_{X^N|Y^N,S^N} = \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k},S^{n-l}}. \tag{3.24}$$

Proof. The proof is similar to the proof of Lemma 3.3.

Proposition 3.6. *For any channel with delay k feedback ($k \geq 1$) and side-information available to both the encoder and decoder with any delay l,*

$$I_{k,l}(X^N \to Y^N S^N) = I(X^N \to Y^N|S^N) \triangleq \sum_{n=1}^{N} I(X^n; Y_n|Y^{n-1}, S^N), \quad \forall N. \tag{3.25}$$

Proof. In the previous subsection, it was shown that the joint distribution of the system could be written as

$$\begin{aligned} P_{X^N,Y^N,S^N} &= \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n} \\ &= P_{S^N} \cdot \vec{P}^{k,l}_{X^N|Y^N,S^N} \cdot \prod_{n=1}^{N} P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{3.26}$$

where $\vec{P}^{k,l}_{X^N|Y^N,S^N}$ is defined in (3.24). Using this in the RHS of Lemma 3.5, we obtain

$$I_{k,l}(X^N \to Y^N S^N) = E\left[\frac{1}{n} \log \frac{\prod_{n=1}^{N} P^{ch}_{Y_n|Y^{n-1},X^n,S^n}}{P_{Y^N|S^N}}\right]. \tag{3.27}$$

In the previous subsection, we also noted that the channel definition was such that

$$P^{ch}_{Y_n|Y^{n-1},X^n,S^N} = P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \quad \forall n.$$

Using this in (3.27), we get

$$\begin{aligned} C^{k,l}_{FB} &= E\left[\frac{1}{n} \log \frac{\prod_{n=1}^{N} P_{Y_n|Y^{n-1},X^n,S^N}}{P_{Y^N|S^N}}\right] \\ &= \frac{1}{N}\left[H(Y^N|S^N) - H(Y^N||X^N|S^N)\right] \\ &= \frac{1}{N} I(X^N \to Y^N|S^N). \end{aligned} \tag{3.28}$$

□

It should be noted that the crucial assumptions needed for the proposition to hold are the property in (3.16) and the fact that the side-information S^N is generated independently. In summary, with k-delayed feedback and l-delayed side-information, the objective function is always the same, viz., $I(X^N \to Y^N|S^N)$. The boost due to feedback and knowledge of side-information the encoder is due the space of optimization varying with k and l. For a channel with feedback delay k and side-information delay l, the space of input distributions is $\mathbf{P}^{k,l}_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$. This space progressively decreases as k and/or l decrease. The general capacity formula for this problem is given by the following theorem.

Theorem 9. *The capacity of a general channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X},\mathbf{S}}$ with k-delayed feedback and side-information available at both encoder and decoder with delay l is*

$$C^{k,l}_{fb} = \sup_{\mathbf{P}^{k,l}_{\mathbf{Y}|\mathbf{X},\mathbf{S}}} \underline{I}(X \to Y|S).$$

Proof. The proof is a generalization of that of Theorem 7 and can be found in Appendix B.2.

3.4 Source-channel separation with feed-forward and feedback

Consider a source $\mathbf{X}$ that has to be transmitted with a distortion D over a channel $\mathbf{W}$ with feedback (feedback delay l). The source has a feed-forward delay k. The source-channel encoder takes a block of N source symbols X^N and maps into channel inputs $A_1, \cdots, A_N$. To produce A_n, the channel input at time n, the source-channel encoder uses X^N as well as the fed-back channel outputs B^{n-l}. To produce the reconstruction $\hat{X}_n$, the source-channel decoder uses the channel outputs B^N and the fed-forward source symbols $X_1, \cdots, X_{n-k}$. The communication set-up is shown in Figure 3.3.

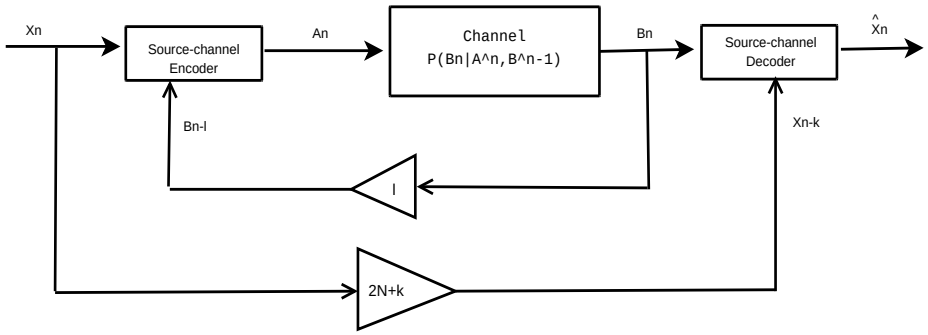

Figure 3.3: Joint source-channel coding system with feedback and feed-forward.

Recall that the source rate-distortion function with feed-forward delay k is given by

$$R_{ff}(D) = \inf \bar{I}_k(\hat{X} \to X) \tag{3.29}$$

where the infimum is over all conditional distributions $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ satisfying the distortion constraint. The feedback capacity of the channel with delay l is given by

$$C_{fb} = \sup \underline{I}(A \to B) \tag{3.30}$$

where the supremum is over all input distributions $\mathbf{P}^l_{\mathbf{X}|\mathbf{Y}}$. Also recall that

$$\bar{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N} \vec{i}_k(\hat{X}^N; X^N) = \limsup_{inprob} \frac{1}{N} \log \frac{P_{X^N,\hat{X}^N}}{\vec{P}^k_{\hat{X}^N|X^N} P_{X^N}} \tag{3.31}$$

$$I_k(\hat{X}^N \to X^N) = E[\log \frac{P_{X^N,\hat{X}^N}}{\vec{P}^k_{\hat{X}^N|X^N} P_{X^N}}] \tag{3.32}$$

and for the channel

$$\underline{I}(A \to B) \triangleq \liminf_{inprob} \frac{1}{N} \vec{i}_l(A^N; B^N) = \liminf_{inprob} \frac{1}{N} \log \frac{P_{A^N,B^N}}{\vec{P}^l_{A^N|B^N} P_{B^N}} \tag{3.33}$$

$$I(A^N \to B^N) = E[\log \frac{P_{A^N,B^N}}{\vec{P}^l_{A^N|B^N} P_{B^N}}] \tag{3.34}$$

The concept of information stability has been studied in detail by Dobrushin, Pinsker and others [20, 62]. On the lines of [81, Definition 3], we give the definition

of *directed information stability* for a source with feed-forward and a channel with feedback. In the remainder of this section, the source and reconstruction alphabets are denoted $\mathcal{X}$ and $\widehat{\mathcal{X}}$, while the channel input and output alphabets are denoted $\mathcal{A}$ and $\mathcal{B}$, respectively.

Definition 3.7. A channel $\mathbf{W} = \{P^{ch}_{B_n|B^{n-1},A^n}\}$ with delay l feedback is called *directed information* stable if there exists an input process $\vec{\mathbf{P}}^{l}_{\mathbf{A}|\mathbf{B}} = \{P_{A_n|A^{n-1},B^{n-l}}\}$ such that the corresponding information density $\vec{i}_l(A^n;B^n)$

$$\frac{\vec{i}_l(A^n;B^n)}{nC_n(W)} \to 1 \text{ in prob },$$

where

$$C_n(W) \triangleq \sup_{\vec{P}^{l}_{A^n|B^n}} \frac{1}{n} I(A^n \to B^n).$$

A source $\mathbf{X} = \{P_{X^n}\}$ with delay k feed-forward is called *directed information* stable if there exists a reconstruction process $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}$ such that the corresponding information density $\vec{i}_k(\hat{X}^n;X^n)$ satisfies

$$\frac{\vec{i}_k(\hat{X}^n;X^n)}{nR_n(D)} \to 1 \text{ in prob },$$

where

$$R_n(D) \triangleq \inf_{P_{\hat{X}^n|X^n}:E[d_n(X^n,\hat{X}^n)\le D]} \frac{1}{n} I_k(\hat{X}^n \to X^n).$$

We note that $R_n(D)$ and $C_n(W)$ are not the same as the feed-forward rate-distortion function $R_{ff}(D)$ and the feedback capacity C_{fb}. In general, $R_n(D)$ and $C_n(W)$ may not have operational significance. We now present a source-channel separation theorem for the system with feedback and feed-forward if the source $\mathbf{X}$ and the channel W are both directed information stable.

Theorem 10. *Let* $\mathbf{X}$ *be a source with feed-forward delay* k *that has to be transmitted over a channel* $\mathbf{W}$ *with feedback delay* l*. If*

1. *The source and channel are both directed information stable,*

2. *The limits* $\lim_{n\to\infty} C_n$ *and* $\lim_{n\to\infty} R_n(D)$ *in Definition 3.7 exist,*

then $\mathbf{X}$ *is transmissible over the channel* $\mathbf{W}$ *if and only if*

$$R_{ff}(D) \le C_{fb}(W).$$

Here, $R_{ff}(D)$ *and* $C_{fb}(W)$ *denote the feed-forward rate-distortion function and feedback capacity, respectively.*

Proof. Direct Part: Suppose $R_{ff}(D) < C_{fb}(W)$, say $R_{ff}(D)+\delta = C_{fb}(W)$. Then by Theorems 2 and 7, we can first build a source code with feed-forward with distortion $\le D$ and rate at most $R_{ff}(D) + \delta/2$. The indices of the source code can be reliably transmitted over the feedback channel since $R_{ff}(D) + \delta/2 < C_{fb}(W)$. This proves the direct part.

Converse: Assume that the source X with feed-forward (with delay k) is transmissible with distortion D over the channel $\{P^{ch}(B_n|A^n, B^{n-1})\}$ with feedback (with delay l) for all sufficiently large blocklengths N using some sequence of joint source-channel codes. The sequence of codes induces a joint process $\{X^N, A^N, B^N, \hat{X}^N\}$ for every N. In the rest of this proof, information quantities are defined using this distribution, unless otherwise specified. We have

$$\begin{aligned} R_N(D) &\overset{(a)}{\le} I_k(\hat{X}^N \to X^N) \\ &= \sum_{n=1}^{N} I(\hat{X}^{n+k-1}; X_n|X^{n-1}) \\ &\overset{(b)}{\le} I(B^N; X_n|X^{n-1}) \\ &= I(B^N; X^N) \\ &= \sum_{n=1}^{N} I(B_n; X^N|B^{n-1}). \end{aligned} \tag{3.35}$$

In the above series of equations, (a) is from the definition of $R_n(D)$. Since the reconstruction at time $n+k-1$ is a function of the channel output sequence B^N and the fed-forward source symbols X^{n-1}, we have the following Markov chain: $X_n - B^N, X^{n-1} - \hat{X}^{n+k-1}$. We obtain (b) by applying the data-processing inequality to this Markov chain. Now,

$$\begin{aligned}
\sum_{t=1}^{N} I(B_n; X^N|B^{n-1}) &= \sum_{n=1}^{N} H(B_n|B^{n-1}) - H(B_n|B^{n-1}, X^N) \\
&\overset{(c)}{=} \sum_{n=1}^{N} H(B_n|B^{n-1}) - H(B_n|B^{n-1}, A^{n+k-1}, X^N) \\
&\overset{(d)}{=} \sum_{n=1}^{N} H(B_n|B^{n-1}) - H(B_n|B^{n-1}, A^n) \\
&= I(A^N \to B^N) \\
&\overset{(e)}{\leq} C_N(W).
\end{aligned} \tag{3.36}$$

(c) is true because at every time instant n, the source-channel encoder produces the channel input A_n as a function of the source sequence X^n and the fed-back channel outputs B^{n-1}. (d) holds because of the channel property- given all the past information up to time n, the channel output B_n depends only on (B^{n-1}, A^n). Note that $A_n, A_{n+1}, \cdots, A_{n+k-1}$ can be considered past information at time n because this can be determined at time n using (B^{n-1}, A^n). (e) follows from the definition of $C_n(W)$. Hence we have for all N,

$$R_N(D) \leq C_N(W). \tag{3.37}$$

Let $P^*_{\hat{X}^N|X^N}$ be the sequence of conditional distributions that achieves the infimum in $R_N(D)$ for each N. We denote this conditional random process $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}} \triangleq \{P^*_{\hat{X}^n|X^N}\}_{N=1}^{\infty}$. Similarly, let $\mathbf{Q}^{l*}_{\mathbf{A}|\mathbf{B}} \triangleq \{\vec{Q}^{l*}_{A^N|B^N}\}_{N=1}^{\infty}$ be the sequence of distributions that achieve $C_N(W)$. For brevity, we will use $\mathbf{P}^*$ and $\mathbf{Q}^*$ to denote the corresponding joint distributions $\mathbf{P}_{\mathbf{X}}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ and $\mathbf{Q}^{l*}_{\mathbf{A}|\mathbf{B}}\mathbf{W}$. Then, from the definition of directed

information stability and due to the fact that $\lim_{N\to\infty} R_N(D)$ exists, we have

$$\liminf_{in\ prob\ P^*} \frac{1}{N}\vec{i}_k(\hat{X}^N;X^N) = \lim_{N\to\infty} R_N(D) = \limsup_{in\ prob\ P^*} \frac{1}{N}\vec{i}_k(\hat{X}^N;X^N). \tag{3.38}$$

Due to similar reasons, we also have

$$\liminf_{in\ prob\ Q^*} \frac{1}{N}\vec{i}_l(A^N;B^N) = \lim_{N\to\infty} C_N(W) = \limsup_{in\ prob\ Q^*} \frac{1}{N}\vec{i}_l(A^N;B^N). \tag{3.39}$$

The two equations above, combined with (3.37) yield

$$\limsup_{in\ prob\ P^*} \frac{1}{N}\vec{i}_k(\hat{X}^N;X^N) \le \liminf_{in\ prob\ Q^*} \frac{1}{N}\vec{i}_l(A^N;B^N). \tag{3.40}$$

Finally, from the formulas for the rate-distortion function and capacity (cf. (3.29) and (3.30)), we have

$$\begin{aligned} R_{ff}(D) &\le \limsup_{in\ prob\ P^*} \frac{1}{N}\vec{i}_k(\hat{X}^N;X^N), \\ C_{fb}(W) &\ge \liminf_{in\ prob\ Q^*} \frac{1}{N}\vec{i}_l(A^N;B^N). \end{aligned} \tag{3.41}$$

Combining this with (3.40), we obtain $R_{ff}(D) \le C_{fb}(W)$. Thus any source with feed-forward that is transmissible over the feedback channel with distortion D has to satisfy

$$R_{ff}(D) \le C_{fb}(W) \tag{3.42}$$

irrespective of the source-channel code used. □

Remark: There have been several papers attempting to establish necessary and sufficient conditions for separation of general channels and sources (without feed-forward or feedback). The source-channel separation theorem above is on the lines of [20, 62]. Our result may be viewed as a generalization of their information stability conditions for source-channel separation to problems with feed-forward and feedback. More general conditions for source-channel separation exist (cf. [81, 36, 37]), and we believe it is possible to extend our separation result with feed-forward and feedback to these general conditions as well.

3.5 Conclusion

In this chapter, we gave an interpretation of the information quantities characterizing the performance limit of channel coding with delayed feedback. This interpretation was then used to find the best achievable rates in a problem where there is some delayed side-information available to both the encoder and decoder, in addition to the feedback or feed-forward. It is worthwhile to explore if the interpretation can help in obtaining/understanding the performance limits of other interesting communication problems. Finally, we established a source-channel separation theorem with feedback and feed-forward under the condition that both the source and channel are directed information stable.

CHAPTER 4

Evaluating the Rate-Distortion Function of Sources with Feed-forward and the Capacity of Channels with Feedback

4.1 Introduction

In the two previous chapters, we characterized the rate-distortion function of sources with (delayed) feed-forward and the capacity of channels with (delayed) feedback. The expressions for the feed-forward rate-distortion function and feedback capacity involve optimization of multi-letter expressions over an infinite dimensional space of distributions (cf. Theorems 2 and 7). These are difficult to compute, and it is not possible to have an efficient algorithm like the Blahut-Arimoto algorithm (cf. [7]) to perform the optimization.

In this chapter, we present a different approach to the problem of computing the rate-distortion and capacity expressions (with feed-forward and feedback). In Sections 4.2 and 4.3, we obtain the structure of the distortion (cost, resp.) function in order for a given joint distribution to achieve the optimum rate-distortion function (channel capacity, resp.). For discrete memoryless channels and sources without feedback/feed-forward, Csiszár and Körner[19] use such an approach to characterize the structure of the cost/distortion function. Our results may be viewed as an extension of the results in [19] to problems with delayed feedback and feed-forward. This

approach is especially relevant since it is otherwise infeasible to calculate the performance limits with feedback and feed-forward. In Section 4.4, we provide examples to illustrate how the structural results can be used to compute many rate-distortion and capacity expressions.

4.2 Evaluating the feed-forward rate-distortion function

The rate-distortion function with k-delay feed-forward, $R_{ff}^{k}(D)$, is the infimum of all achievable rates at distortion D with k-delay feed-forward. This was characterized in Chapter 2 and we restate the result below. We first recall definitions of some required quantities.

$$\vec{P}^{k}_{\hat{X}^N|X^N} = \prod_{n=1}^{N} P_{\hat{X}_n|X^n,\hat{X}^{n-1}} \tag{4.1}$$

$$\bar{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N} \vec{i}_k(\hat{X}^N \to X^N) = \limsup_{inprob} \frac{1}{N} \log \frac{P_{X^N,\hat{X}^N}}{\vec{P}^{k}_{\hat{X}^N|X^N} P_{X^N}} \tag{4.2}$$

$$I_k(\hat{X}^N \to X^N) = E[\log \frac{P_{X^N,\hat{X}^N}}{\vec{P}^{k}_{\hat{X}^N|X^N} P_{X^N}}] \tag{4.3}$$

For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with k-delay feed-forward is given by

$$R_{ff}^{k}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}_k(\hat{X} \to X), \tag{4.4}$$

where $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ and

(4.5)

$$\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) = \inf \left\{ h : \lim_{n\to\infty} P_{X^n,\hat{X}^n}\left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h\right) = 0 \right\}$$

This is an optimization over an infinite dimensional space of conditional distributions $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$. Since this is a potentially difficult optimization, we can turn the problem on its head and pose the following question:

Given a source X with distribution $\mathbf{P_X}$ and a conditional distribution $\mathbf{P_{\hat{X}|X}}$, for what sequence of distortion measures does $\mathbf{P_{\hat{X}|X}}$ achieve the infimum in the rate-distortion formula ?

A similar approach is used in [19] (Problem 2 and 3, p. 147) to find optimizing distributions for discrete memoryless channels and sources without feedback/feed-forward. It is also used in [29] to study the optimality of transmitting uncoded source data over channels and in [66] to study the duality between source and channel coding.

We now present a theorem that gives conditions for a given joint process to achieve the optimum in the rate-distortion formula. The joint process is assumed to be *directed information stable*. The concept information stability of random processes is studied in great detail in the book by Pinsker [62]. In [78], the concept is extended to directed information stability.

Definition 4.1. A joint process $\mathbf{P_{X\hat{X}}}$ is k-directed information stable if

$$\lim_{N\to\infty} P\left(\left|\frac{\vec{i}_k(\hat{X}^N \to X^N)}{I_k(\hat{X}^N \to X^N)} - 1\right| > \epsilon\right) = 0 \quad \forall \epsilon > 0.$$

For any joint process $\mathbf{P_{X\hat{X}}}$, it is true [78] that

$$\underline{I}_k(\hat{X} \to X) \le \liminf_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N) \le \limsup_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N) \le \bar{I}_k(\hat{X} \to X.)$$

If a joint process $\mathbf{P_{\hat{X}X}}$ is k-directed information stable, it can be shown (cf. [78]) that

(4.6)

$$\underline{I}_k(\hat{X} \to X) = \liminf_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N) \le \limsup_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N) = \bar{I}_k(\hat{X} \to X).$$

We point out that information stability of a random process is related to information stability of a source/channel introduced in the previous chapter. It is possible to

define a directed information stable source or channel in terms of directed information stable processes.

Given a source $\mathbf{X}$, consider a joint process that is k-directed information stable. If for the process, the lim sup and lim inf in (4.6) are equal (i.e., the limit exists), the following theorem characterizes the distortion measures for which the joint process achieves the optimum.

Theorem 11. *Suppose we are given a source X characterized by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$ with feed-forward delay k and $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ is a conditional distribution such that the joint process is k-directed information stable and equality holds in (4.6). Then $\mathbf{P_{\hat{X}|X}}$ achieves the rate-distortion function with k-delay feed-forward at distortion level D if for all sufficiently large n, the distortion measure satisfies*

$$d_n(x^n, \hat{x}^n) = -c \cdot \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}(x^n, \hat{x}^n)}{\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n)} + d_0(x^n), \tag{4.7}$$

where

$$\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \prod_{i=1}^{n} P_{\hat{X}_i|X^{i-k},\hat{X}^{i-1}}(\hat{x}_i|x^{i-k}, \hat{x}^{i-1})$$

and c is any positive number, $d_0(.)$ is an arbitrary function, and

$$D = \limsup_{n\to\infty} E d_n(X^n, \hat{X}^n)$$

.

Proof. See Appendix C.1. □

Our theorem requires the chosen joint process to be directed information stable and satisfy (4.6) with equality. We remark that this is a fairly weak condition that includes stationary, ergodic processes as well as processes that are asymptotically stationary. The chosen conditional distribution process is such that the joint process

satisfies the stability conditions, but we emphasize that the theorem gives the condition for optimality of $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ among *all* valid conditional distributions, not just the ones that make the joint process information stable.

The formulation relating the structure of the distortion function to the joint distribution first appears in [19] - for discrete memoryless sources and channels with no feedback/feed-forward and single-letter additive distortion/cost measures. The theorems in this chapter extend the structural results of [19] to general sources (and channels) and distortion(cost) measures in the presence of feed-forward(feedback). This is interesting because feedback/feed-forward necessitates the use of information spectrum methods which makes it infeasible to directly compute the optimization.

We note that the results in [29] and [66] are different since they directly apply the structural results of [19] to: a) study the optimality of uncoded transmission of sources over channels [29], and b) study duality between source and channel coding [66].

4.2.1 Markov sources with feed-forward

Markov sources, an important class of sources with memory. A stationary, ergodic mth order Markov source $\mathbf{X}$ is characterized by a distribution $\mathbf{P}_{\mathbf{X}} = \{P_{X^n}\}_{n=1}^{\infty}$ where

$$P_{X^n} = \prod_{i=1}^{n} P_{X_i|X_{i-m}^{i-1}}, \quad \forall n \tag{4.8}$$

and the $P_{X_i|X_{i-m}^{i-1}}$ since $\mathbf{X}$ is stationary. Let the system have feed-forward with delay k. We ask: *When is the optimal joint distribution mth order Markov in the following sense*:

$$P_{X^n,\hat{X}^n} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i|X_{i-m}^{i-1}}, \quad \forall n. \tag{4.9}$$

In other words, when does the optimizing conditional distribution to have the form

$$P_{\hat{X}^n|X^n} = \prod_{i=1}^{n} P_{\hat{X}_i|X_{i-m}^i}, \quad \forall n. \tag{4.10}$$

This question can be answered using Theorem 11 and is stated as a corollary below. In the rest of this subsection, we drop the subscripts on the probabilities to keep the notation clean.

Corollary 4.2. *For an mth order Markov source (as described in* (4.8)*) with feed-forward delay k, an mth order conditional distribution (as described in* (4.10)*) achieves the optimum in the rate-distortion function with k-delay feed-forward at distortion level D for a sequence of distortion measures* $\{d_n\}$ *given by*

$$d_n(x^n, \hat{x}^n) = -c \cdot \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(x_i, \hat{x}_i|x_{i-m}^{i-1})}{P(\hat{x}_i|\hat{x}_{i-k+1}^{i-1}, x_{i-k+1-m}^{i-k})} + d_0(x^n), \tag{4.11}$$

where c is any positive number and $d_0(.)$ *is an arbitrary function, and*

$$D = \limsup_{n\to\infty} E d_n(X^n, \hat{X}^n)$$

.

Proof. See Appendix C.2.

In Section 4.4, we will discuss examples that illustrate the use of Corollary 4.2.

4.3 Evaluating the channel capacity with feedback

In this section, we consider channels with feedback and the problem of evaluating their capacity. The channel is defined as a sequence of probability distributions:

$$P_{\mathbf{Y}|\mathbf{X}}^{ch} = \{P_{Y_n|X^n,Y^{n-1}}^{ch}\}_{n=1}^{\infty} \tag{4.12}$$

where X_n and Y_n are the channel input and output symbols at time n, respectively. The channel is assumed to have $k-$delay feedback ($1 \leq k < \infty$). The input distribution to the channel is denoted by $P_{\mathbf{X}|\mathbf{Y}}^{k} = \{P_{X_n|X^{n-1},Y^{n-k}}\}_{n=1}^{\infty}$.The joint distribution

of the system is given by $P_{\mathbf{X},\mathbf{Y}} = \{P_{X^n,Y^n}\}_{n=1}^{\infty}$, where

$$\vec{P}^{ch}_{Y^n|X^n} \triangleq \prod_{i=1}^{n} P_{Y_i|X^i,Y^{i-1}}, \vec{P}^{k}_{X^n|Y^n} \quad \triangleq \prod_{i=1}^{n} P_{X_i|X^{i-1},Y^{i-k}} \tag{4.13}$$

$$P_{X^n,Y^n} = \vec{P}^{k}_{X^n|Y^n} \cdot \vec{P}^{ch}_{Y^n|X^n}$$

with

Thus far we considered feedback channels without a cost constraint. Let us now introduce a cost function associated with using the channel. Let $c_N(X^N, Y^N)$ be the cost for N uses of the channel. For example, this could be the average power of the input symbols. Note that in general, we have allowed the cost function at time N to depend on the inputs and the outputs until time N. This is because the encoder learns the outputs (with some delay) due to the feedback. So it can potentially use this information to choose future input symbols so that the cost constraint is satisfied. In the case of no feedback, the dependence of the cost function on the channel output can be averaged out, resulting in a new cost function that depends only on the channel input.

The probability of error is defined in the usual way. If W is the message (with uniform distribution over a finite set) that was transmitted, then

$$P_e = Pr(g(Y^N) \neq W).$$

We now define an achievable rate with $k-$delay feedback at cost P.

Definition 4.3. R is an (ϵ, δ)-achievable rate at cost P with k-delay feedback if for all sufficiently large N, there exists an $(N, 2^{NR})$ channel code such that

$$P_e < \epsilon,$$

$$Pr(c_N(X^N, Y^N) > P) < \delta$$

R is an achievable rate at cost P with k-delay feedback if it is (ϵ, δ)-achievable for every $\epsilon, \delta > 0$.

The feedback capacity results in [77, 78] can be extended to include a cost constraint and we state this as a fact below.

Fact. For an arbitrary channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$, the capacity with k-delay feedback at cost P is given by

$$C^k_{fb}(P) = \sup_{\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}:\rho(P^k_{\mathbf{X}|\mathbf{Y}})\leq P} \underline{I}(X \to Y), \tag{4.14}$$

where

(4.15)

$$\rho(\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}) \triangleq \limsup_{inprob} c_n(X^n, Y^n) = \inf\left\{h : \lim_{n\to\infty} P_{X^nY^n}\left((x^n, y^n) : c_n(x^n, y^n) > h\right)\right\} = 0.$$

The capacity formula above is a multi-letter expression involving optimizing the function

$$\underline{I}(X \to Y) \triangleq \liminf_{inprob} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}$$

over an infinite dimensional space of input distributions $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$. Just like we did with sources, we can pose the following question: *Given a channel* $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ *and an input distribution* $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$*, for what sequence of cost measures does* $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$ *achieve the supremum in the capacity formula ?*

Consider a joint process $\mathbf{P}_{\mathbf{XY}} = \mathbf{P}^k_{X|Y}\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ is directed information stable. Recall from the discussion in 3.2.1 that due to the constraint on the input distribution in delay k feedback, the k-directed information is equal to the directed information for all k. Thus Definition 4.1 is the same for feedback with any delay $k \geq 1$, and we refer to it as just directed information stability. As remarked in the previous section,

directed information stability of $\mathbf{P_{XY}}$ implies

(4.16)

$$\underline{I}(X \to Y) = \liminf_{N\to\infty} \frac{1}{N} I(X^N \to Y^N) \leq \limsup_{N\to\infty} \frac{1}{N} I(X^N \to Y^N) = \bar{I}(X \to Y).$$

If the lim sup and lim inf in the above equation are the same (i.e., the limit exists), the following theorem gives the structure of the cost function for the process $\mathbf{P_{XY}}$ to achieve the optimum.

Theorem 12. *For a channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback, let $\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}}$ be an input distribution such that the joint process $\mathbf{P}_{\mathbf{X},\mathbf{Y}} = \{P_{X^n,Y^n}\}_{n=1}^{\infty}$ (given by (4.13)) is directed information stable and (4.16) holds with equality. Then the input distribution $P^{k}_{\mathbf{X}|\mathbf{Y}}$ achieves the k-delay feedback capacity of the channel at cost level P if for all sufficiently large n, the cost measure satisfies*

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)} + d_0, \tag{4.17}$$

where $\vec{P}^{ch}_{Y^n|X^n}$ is defined in (4.13), λ is any positive number, d_0 is an arbitrary constant and $P = \limsup_{n\to\infty} c_n(x^n, y^n)$.

Proof. See Appendix C.3. □

In the theorem, we have considered an input process $\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}}$ such that the joint process $\mathbf{P}_{\mathbf{X},\mathbf{Y}}$ is information stable and satisfies (4.16) with equality. This is a fairly weak condition and includes stationary as well as asymptotically stationary processes. The chosen input distribution is such that the joint process satisfies the stability conditions, but we emphasize that the theorem gives the condition for optimality of $\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}}$ among *all* valid input distributions, not just the ones that make the joint process information stable.

4.3.1 Markov channels with feedback

We consider now a simple Markov channel with feedback and the problem of evaluating its capacity. The Markov channel we study is characterized by

$$P^{ch}_{Y_n|X^n,Y^{n-1}} = P^{ch}_{Y_n|X_n,Y_{n-1}}. \tag{4.18}$$

Let the channel have feedback with delay 1. The problem of evaluating the capacity of finite state machine channels was studied in [91] and [12]. In [78], it was shown that the capacity of such a channel is achieved by a feedback dependent Markov input distribution. This means that the input distribution $\{P_{X_n|X^{n-1},Y^{n-1}}\}$ is Markov in X but depends on all the past Y symbols. It was also shown that the problem of finding the capacity of a finite state machine channel with feedback could be formulated as a stochastic control problem that could be solved using dynamic programming in certain cases.

We are interested in finding cost measures for which the capacity of the channel in (4.18) is easily evaluated. We first ask: *When is the optimal joint distribution first order Markov in the following sense*:

$$P_{X^n,Y^n} = \prod_{i=1}^{n} P_{X_i,Y_i|Y_{i-1}}, \quad \forall n. \tag{4.19}$$

In other words, when does the optimizing input distribution to have the form

$$\vec{P}_{X^n|Y^n} = \prod_{i=1}^{n} P_{X_i|Y_{i-1}}, \quad \forall n. \tag{4.20}$$

From Theorem 12, it is seen that this happens when the cost-function has the form:

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \sum_{i=1}^{n} \log \frac{P^{ch}_{Y_i|X_i,Y_{i-1}}(y_i|x_i, y_{i-1})}{P_{Y_i|Y_{i-1}}(y_i|y_{i-1})} + d_0. \tag{4.21}$$

This can be shown as follows. From (4.19), we see that the marginal distribution of Y^n has the form

$$P_{Y^n} = \sum_{X^n} P_{X^n,Y^n} = \sum_{X^n} \prod_{i=1}^{n} P_{X_i,Y_i|Y_{i-1}} = \prod_{i=1}^{n} P_{Y_i|Y_{i-1}}. \tag{4.22}$$

Substituting (4.18) and (4.22) in Theorem 12, we obtain the structure of the cost function given by (4.21). In the next section, we will present an example of a Markov channel as described in (4.18) and evaluate its feedback capacity-cost function. Clearly, other kinds of Markov channels and input distributions can be considered too; Theorem 12 then gives the structure of the cost function for optimality.

4.4 Examples

We now provide a few examples to illustrate how Theorem 11 and 12 can be used to determine the rate-distortion function (capacity) of sources (channels) with feed-forward (feedback).

4.4.1 Source coding examples

In our first example, we consider an asymmetric binary Markov source and obtain its rate-distortion function with feed-forward. Our next example deals with a stock price variation problem- we model the problem using a finite state Markov chain with feed-forward. In our third example, we will consider a Gauss-Markov source with feed-forward. We will obtain the results using Theorem 11.

In [85], the optimal distortion-rate function with feed-forward was derived for a class of sources (those that can be represented auto-regressively with an innovations process, where the innovations process is either IID or satisfies the Shannon Lower Bound (SLB)[15] with equality). Using this result, the distortion-rate function for a symmetric binary Markov source with feed-forward and a stationary Gaussian source with feed-forward were evaluated. We remark that in our first two examples, the innovations processes are *not* IID and the feed-forward rate-distortion functions cannot be computed using the results in [85].

Table 4.1: Distortion $e\left(\hat{x}_i, x_{i-1}, x_i\right)$

	(x_{i-1}, x_i)			
	00	01	10	11
$\hat{x}_i = 0$	0	0	0	1
$\hat{x}_i = 1$	1	1	1	0

Binary Asymmetric Markov Source

We now consider a binary asymmetric first-order Markov source with both $\mathcal{X}$ and $\hat{\mathcal{X}}$ equal to $\{0, 1\}$. The source has the following transition probability matrix:

$$\begin{aligned} P(X_i = 0|X_{i-1} = 1) = q \qquad & P(X_i = 1|X_{i-1} = 0) = p \\ P(X_i = 0|X_{i-1} = 0) = 1 - p \qquad & P(X_i = 1|X_{i-1} = 1) = 1 - q, \quad \forall i \end{aligned} \tag{4.23}$$

Suppose the decoder needs to detect all consecutive occurrences of 1 in the source. To achieve this, we have a finite rate R as well as feed-forward. An error occurs either when the decoder fails to detect a 11 pattern or when it falsely detects a 11 pattern that did not occur. We use a Hamming distortion criterion

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} e(\hat{x}_i, x_{i-1}, x_i), \tag{4.24}$$

where $e(.,.,.)$ is the *per-letter* distortion given Table 4.1. This example is inspired by the one given in [27, Sec. 9.8, p. 490].

Proposition 4.4. *For a binary first-order Markov source described by* (4.23) *and a distortion function given by* (4.24) *and Table 4.1, the rate distortion function with delay* 1 *feed-forward is given by*

$$R_{ff}(D) = \begin{cases} \frac{p}{p+q}[h(q) - h(D(1 + \frac{q}{p}))] & \text{if } D < \frac{p}{p+q} \min\{q, 1-q\}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.25}$$

where $h(.)$ is the binary entropy function.

Proof. The stationary distribution of the Markov chain is

$$[\pi_0 \qquad \pi_1] = \left[\frac{q}{p+q} \qquad \frac{p}{p+q}\right].$$

We first note that $R_{ff}(D)$=0 for $D \geq \pi_1 \min\{q, 1-q\}$. This is because the decoder knows X_{n-1} due to feed-forward. It always declares $\hat{X}_n = 0$ when $X_{n-1} = 0$. When $X_{n-1} = 1$, if the decoder always declares $\hat{X}_n = 0$, we make an error with probability $1-q$. On the other hand, if the decoder always declares $\hat{X}_n = 1$ when $X_{n-1} = 1$, we make an error with probability q.

We now consider the case when $D < \pi_1 \min\{q, 1-q\}$. As in the previous section, we will use Corollary 4.2 to verify that a binary first-order Markov conditional distribution of the form

$$P_{\hat{X}_n|\hat{X}^{n-1},X^n} = P_{\hat{X}_n|X_n,X_{n-1}}, \quad \forall n \tag{4.26}$$

achieves the optimum (this condition is the same as (4.10)).

Due to the structure and symmetry of the distortion function in Table 4.1, we further choose $P(x_n|\hat{x}_n, x_{n-1})$ to have the structure shown in Table 4.2. The intuition behind this is as follows. When $X_{n-1} = 0$, the decoder can always declare $\hat{X}_n = 0$, there is no error irrespective of the value of X_n. So we assign $P(\hat{X}_n = 0|x_{n-1} = 0, x_n = i) = 1$, for $i = 0, 1$. This gives

$$P(X_n = 0|x_{n-1} = 0, \hat{x}_n = 0) = 1 - p.$$

The event $(X_{n-1} = 0, \hat{X}_n = 1)$ has zero probability. When $X_{n-1} = 1$ and $\hat{X}_n = 0$, there is no error if $X_n = 0$ and an error occurs if $X_n = 1$. Hence $P(X_n = 1|x_{n-1} = 1, \hat{x}_n = 0)$ is assigned a value ϵ, where ϵ will be determined using the distortion constraint. Due to symmetry, the case $X_{n-1} = 1, \hat{X}_n = 1$ is treated in the same way. For this distribution, we can show that the distortion criterion (4.24) can be cast in the following form

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} \Big(- c \log_2 P(x_i|\hat{x}_i, x_{i-1}) + d_0(x_{i-1}, x_i) \Big), \tag{4.27}$$

Table 4.2: The distribution $P(x_i|x_{i-1},\hat{x}_i)$

$(x_{n-1},\hat{x}_n)$	00	01	10	11
$x_n=0$	$1-p$	$-$	$1-\epsilon$	ϵ
$x_n=1$	p	$-$	ϵ	$1-\epsilon$

Table 4.3: The conditional distribution $P(\hat{x}_i|x_{i-1},x_i)$

(x_{i-1},x_i)	00	01	10	11
$\hat{x}_i=0$	1	1	$\frac{(q-\epsilon)(1-\epsilon)}{q(1-2\epsilon)}$	$\frac{\epsilon(q-\epsilon)}{(1-q)(1-2\epsilon)}$
$\hat{x}_i=1$	0	0	$\frac{\epsilon(1-q-\epsilon)}{q(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q-\epsilon)}{(1-q)(1-2\epsilon)}$

or equivalently

$$e(\hat{x}_i,x_{i-1},x_i) = -c\log_2 P(x_i|\hat{x}_i,x_{i-1}) + d_0(x_{i-1},x_i), \tag{4.28}$$

proving that the distribution in Table 4.2 is optimal. c, $d_0(0,0)$, $d_0(0,1)$, $d_0(1,1)$ and $d_0(1,0)$ are determined by substituting values from Tables 4.1 and 4.2 into (4.28):

$$c = \frac{1}{\log(1-\epsilon)-\log\epsilon}, \quad d_0(0,0) = c\log(1-p), \quad d_0(0,1) = c\log p,$$

$$d_0(1,0) = c\log(1-\epsilon), \quad d_0(1,1) = c\log(1-\epsilon).$$

The process $\{\mathbf{X},\hat{\mathbf{X}}\} = \{X^n,\hat{X}^n\}_{n=1}^{\infty}$ is stationary and ergodic. When $(x^n,\hat{x}^n) \sim P_{X^n,\hat{X}^n}$, by the law of large numbers the distortion

$$\begin{aligned} d_n(x^n,\hat{x}^n) &= \frac{1}{n}\sum_{i=1}^{n} e(\hat{x}_i,x_{i-1},x_i) \\ &\to E[e(\hat{x}_i,x_{i-1},x_i)] \quad \text{as} \quad n\to\infty \quad w.p.1. \end{aligned}$$

Therefore the distortion constraint is equivalent to $E[e(\hat{x}_2,x_1,x_2)] \le D$. To calculate the expected distortion

$$E[e(\hat{x}_2,x_1,x_2)] = \sum_{x_1,x_2,\hat{x}2} P(x_1,x_2)P(\hat{x}_2|x_1,x_2)e(\hat{x}_2,x_1,x_2), \tag{4.29}$$

we need to compute the (optimum achieving) conditional distribution $P(\hat{x}_2|x_1,x_2)$. This is done by using the relation

$$P(x_2|x_1,\hat{x}_2) = \frac{P(x_2|x_1)P(\hat{x}_2|x_2,x_1)}{\sum_{x_2} P(x_2|x_1)P(\hat{x}_2|x_2,x_1)}. \tag{4.30}$$

In the above equation, we can solve for the values of $P(\hat{x}_2|x_1, x_2)$ by substituting the values from Table 4.2 for $P(x_2|x_1, \hat{x}_2)$ and from (4.23) for $P(x_2|x_1)$. Thus we obtain the conditional distribution $P(\hat{x}_2|x_1, x_2)$ shown in Table 4.3. Using this in (4.29), we obtain

$$E[e(\hat{x}_2, x_1, x_2)] = \epsilon\pi_1 \leq D \tag{4.31}$$

We can now calculate the rate distortion function as

$$\begin{aligned} R^1_{ff}(D) = \overline{I}(\hat{X} \to X) &= \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N, \hat{x}^N} P(x^N, \hat{x}^N) \log_2 \frac{\prod_{n=1}^N P(x_n|\hat{x}^n, x^{n-1})}{\prod_{n=1}^N P(x_n|x^{n-1})} \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N, \hat{x}^N} P(x^N, \hat{x}^N) \log_2 \prod_{n=1}^N \frac{P(x_n|\hat{x}_n, x_{n-1})}{P(x_n|x_{n-1})} \\ &= \sum_{x_1, x_2, \hat{x}_2} P(x_1, x_2, \hat{x}_2) \log_2 \frac{P(x_2|x_1, \hat{x}_2)}{P(x_2|x_1)} \\ &= \pi_1 \left(h(q) - h(\epsilon)\right) = \pi_1 \left(h(q) - h(\frac{D}{\pi_1})\right). \end{aligned} \tag{4.32}$$

□

Proposition 4.4 has a nice interpretation in terms of the innovations process. Note that there is a possibility of the decoder making an error at time i only when $X_{i-1} = 1$. So, only innovations at time instants i when $X_{i-1} = 1$ need to be conveyed to the decoder. Since $P(X_i = 1) = \pi_1$, this means only a fraction π_1 of the innovations have to be encoded. Also, to achieve an overall distortion D, we can afford to encode these innovations with distortion $\frac{D}{\pi_1}$.

We recall that the 'no feed-forward' rate-distortion function to reconstruct an i.i.d binary process $X \sim Bernoulli(q)$ with Hamming distortion D/π_1 is $h(q) - h(D/\pi_1)$. The overall rate in bits/sample is π_1 times this value since we encode only a fraction π_1 of the innovations.

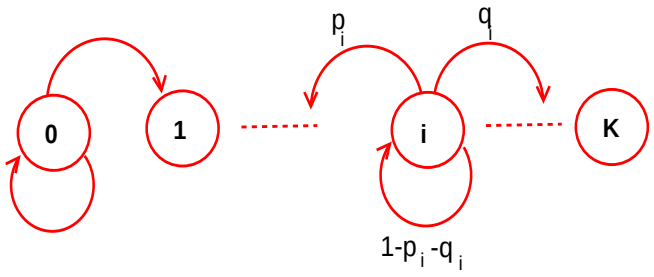

Figure 4.1: Markov chain representing the stock value

Stock-market example

Suppose that we wish to observe the behavior of a particular stock in the stock market over an $N-$day period. Assume that the value of the stock can take $k+1$ different values and is modeled as a $k+1$-state Markov chain, as shown in Fig. 4.1. If on a particular day, the stock is in state i, $1 \leq i < k$, then on the next day, one of the following can happen.

- The value increases to state $i+1$ with probability p_i.
- The value drops to state $i-1$ with probability q_i.
- The value remains the same with probability $1-p_i-q_i$.

When the stock-value is in state 0, the value cannot decrease. Similarly, when in state k, the value cannot increase. Suppose an investor invests in this stock over an $N-$day period and desires to be forewarned whenever the value drops. We assume that there is an insider (who has a priori information about the behavior of the stock over the N days) who can send information to the investor at a finite rate.

The value of the stock is modeled as a Markov source $\mathbf{X} = \{X_n\}$. The reconstruction $\hat{X}_n$ is binary: $\hat{X}_n = 1$ indicates that the price is going to drop from day $n-1$ to n, $\hat{X}_n = 0$ means otherwise. Before day n, the investor knows all the previous values of the stock X^{n-1} and has to make the decision $\hat{X}_n$. Thus feed-forward is

Table 4.4: Distortion $e\left(\hat{x}_i, x_{i-1} = j, x_i\right)$

(x_{i-1}, x_i)	$j, j+1$	j, j	$j, j-1$
$\hat{x}_i = 0$	0	0	1
$\hat{x}_i = 1$	1	1	0

Table 4.5: The distribution $P\left(X_i | x_{i-1}, \hat{x}_i\right)$

$(x_{i-1}, \hat{x}_i)$	00	01	$\cdots$	$j0$	$j1$	$\cdots$	$k0$	$k1$
$x_i = 0$	$1-p$	–	$\cdots$	–	–	–	–	–
$x_i = 1$	p	–	$\cdots$	–	–	–	–	–
$x_i = \vdots$	–	–	$\ddots$	–	–	–	–	–
$x_i = j-1$	–	–	–	ϵ	$1-\epsilon$	–	–	–
$x_i = j$	–	–	–	$\frac{(1-\epsilon)(1-p_j-q_j)}{1-q_j}$	$\frac{\epsilon(1-p_j-q_j)}{1-q_j}$	–	–	–
$x_i = j+1$	–	–	–	$\frac{(1-\epsilon)p_j}{1-q_j}$	$\frac{\epsilon p_j}{1-q_j}$	–	–	–
$x_i = \vdots$	–	–	–	–	–	$\ddots$	–	–
$x_i = k-1$	–	–	$\cdots$	–	–	–	ϵ	$1-\epsilon$
$x_i = k$	–	–	$\cdots$	–	–	–	$1-\epsilon$	ϵ

automatically built into the problem.

The investor makes an error either when she fails to predict a drop or when she falsely predicts a drop. The distortion is modeled using a Hamming distortion criterion as follows.

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} e(\hat{x}_i, x_{i-1}, x_i), \tag{4.33}$$

where $e(.,.,.)$ is the *per-letter* distortion given Table 4.4. The minimum amount of information (in bits/sample) the insider needs to convey to the investor so that he can predict drops in value with distortion D is given by $R_{ff}^1(D)$.

Proposition 4.5. *For the stock-market problem described above,*

$$R_{ff}(D) = \sum_{i=1}^{k-1} \pi_i \left(h(p_i, q_i, 1-p_i-q_i) - h(\epsilon, 1-\epsilon) \right) + \pi_k \left(h(q_k, 1-q_k) - h(\epsilon, 1-\epsilon) \right),$$

where $h()$ is the entropy function, $[\pi_0, \pi_1, \cdots, \pi_k]$ is the stationary distribution of the Markov chain and $\epsilon = \frac{D}{1-\pi_0}$.

Table 4.6: The conditional distribution $P(\hat{X}_i|x_{i-1}, x_i)$

	(x_{i-1}, x_i)						
	$0,0$	$0,1$	$j, j-1$	j, j	$j, j+1$	$k, k-1$	k, k
$\hat{x}_i = 0$	1	1	$\frac{\epsilon(1-q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{\epsilon(1-q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$
$\hat{x}_i = 1$	0	0	$\frac{(1-\epsilon)(q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{(1-\epsilon)(q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$

Proof. As before, we will use Corollary 4.2 to verify that a first-order Markov conditional distribution of the form

$$P_{\hat{X}_n|\hat{X}^{n-1},X^n} = P_{\hat{X}_n|X_n,X_{n-1}}, \quad \forall n \tag{4.34}$$

achieves the optimum.

Due to the structure of the distortion function in Table 4.4, we choose the structure of $P(x_i|\hat{x}_i, x_{i-1})$ as follows. When $X_{i-1} = 0$, the decoder can always declare $\hat{X}_i = 0$, there is no error irrespective of the value of X_i. So we assign $P(\hat{X}_i = 0|x_{i-1} = 0, x_i = 0) = 1$, which gives

$$P(X_i = 0|x_{i-1} = 0, \hat{x}_i = 0) = 1 - p.$$

The event $(X_{i-1} = 0, \hat{X}_i = 1)$ has zero probability. Thus we obtain the first two columns of Table 4.5. When $(X_{i-1} = j, \hat{X}_i = 0)$, $1 \leq j \leq k$, an error occurs when $X_i = j-1$. This is assigned a probability ϵ. The remaining probability $(1-\epsilon)$ is split between $P(X_i = j|x_{i-1} = j, \hat{x}_i = 0)$ and $P(X_i = j+1|x_{i-1} = j, \hat{x}_i = 0)$ according to their transition probabilities. In a similar fashion, we obtain all the columns in Table 4.5.

We now show that the distortion criterion (4.33) can be cast in the form

$$d_n(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n}\Big(- c\log_2 P(x_i|\hat{x}_i, x_{i-1}) + d_0(x_{i-1}, x_i)\Big), \tag{4.35}$$

or equivalently

$$e(\hat{x}_i, x_{i-1}, x_i) = -c\log_2 P(x_i|\hat{x}_i, x_{i-1}) + d_0(x_{i-1}, x_i), \tag{4.36}$$

thereby proving that the distribution in Table 4.5 is optimal. This is done by determining the values of $c, d_0(x_{i-1}, x_i)$, $1 \le x_{i-1}, x_i \le k$. Using the values Tables 4.4 and 4.5 in (4.36), we find $c, d_0(.,.)$:

$$\begin{aligned}
c &= \frac{1}{\log(1-\epsilon) - \log\epsilon}, \\
d_0(0,0) &= c\log(1-p), \\
d_0(0,1) &= c\log p, \\
d_0(j,j-1) &= c\log(1-\epsilon), \qquad 1 \le j \le k \\
d_0(j,j) &= c\log\frac{(1-\epsilon)(1-p_j-q_j)}{1-q_j}, \qquad 1 \le j \le k \\
d_0(j,j+1) &= c\log\frac{(1-\epsilon)p_j}{1-q_j}, \qquad 1 \le j \le k-1.
\end{aligned}$$

Since the process $\{\mathbf{X}, \hat{\mathbf{X}}\}$ is jointly stationary and ergodic, when $(x^n, \hat{x}^n) \sim P_{X^n,\hat{X}^n}$, the distortion

$$d_n(x^n, \hat{x}^n) \to E[e(\hat{x}_2, x_1, x_1)] \quad \text{as} \quad n \to \infty \quad w.p.1.$$

Hence the distortion constraint is equivalent to $E[e(\hat{x}_2, x_1, x_2)] \le D$. To calculate the expected distortion

$$E[e(\hat{x}_2, x_1, x_2)] = \sum_{x_1, x_2, \hat{x}2} P(x_1, x_2)P(\hat{x}_2|x_1, x_2) \cdot e(\hat{x}_2, x_1, x_2), \tag{4.37}$$

we need the (optimum achieving) conditional distribution $P(\hat{X}_2|x_1, x_2)$. This is found by substituting the values from Table 4.5 in the relation

$$P(x_2|x_1, \hat{x}_2) = \frac{P(x_2|x_1)P(\hat{x}_2|x_2, x_1)}{\sum_{x_2} P(x_2|x_1)P(\hat{x}_2|x_2, x_1)}. \tag{4.38}$$

Thus we obtain the conditional distribution $P(\hat{X}_2|x_1, x_2)$ shown in Table 4.4.1. Using this in (4.37), we get

$$E[e(\hat{x}_2, x_1, x_2)] = (1 - \pi_0)\epsilon \le D \tag{4.39}$$

We can now calculate the rate distortion function as

$$\begin{aligned} R_{ff}(D) = &= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N,\hat{x}^N} P(x^N,\hat{x}^N) \log_2 \prod_{n=1}^{N} \frac{P(x_n|\hat{x}^n, x^{n-1})}{P(x_n|x^{n-1})} \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N,\hat{x}^N} P(x^N,\hat{x}^N) \log_2 \prod_{n=1}^{N} \frac{P(x_n|\hat{x}_n, x_{n-1})}{P(x_n|x_{n-1})} \\ &= \sum_{x_1,x_2,\hat{x}_2} P(x_1,x_2,\hat{x}_2) \log_2 \frac{P(x_2|x_1,\hat{x}_2)}{P(x_2|x_1)} \\ &= H(X_2|X_1) - H(X_2|\hat{X}_2, X_1) \end{aligned} \tag{4.40}$$

to obtain the expression in Proposition 4.5. □

First Order Gauss-Markov Source

Consider a stationary, ergodic, first-order Gauss-Markov Source X with mean 0, correlation ρ and variance σ^2 described by

$$X_n = \rho X_{n-1} + N_n, \quad \forall n, \tag{4.41}$$

where $\{N_n\}$ are independent, identically distributed Gaussian random variables with mean 0 and variance $(1-\rho^2)\sigma^2$. Assume the source has feed-forward with delay 1. Suppose we want to reconstruct at every time instant n the linear combination $aX_n + bX_{n-1}$ for any constants a, b. We use the mean-squared error distortion criterion:

$$d_n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{x}_i - (ax_i + bx_{i-1})\right)^2. \tag{4.42}$$

Proposition 4.6. *For the stationary first-order Gauss-Markov source described in* (4.41) *and the distortion criterion in* (4.42)*, the optimal rate-distortion function (with feed-forward delay* 1*) - the smallest rate for which* $d_n(X^n,\hat{X}^n) \le D$ *w.p.*1*- is given by*

$$R_{ff}(D) = \frac{1}{2}\log\frac{\sigma^2(1-\rho^2)}{D/a^2}. \tag{4.43}$$

Proof. We will use Corollary 4.2 to verify that a Gaussian first-order Markov conditional distribution achieves the optimum. A first-order Markov conditional distribution $\{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ satisfies

$$P_{\hat{X}_n|\hat{X}^{n-1},X^n} = P_{\hat{X}_n|X_n,X_{n-1}}, \quad \forall n. \tag{4.44}$$

For this distribution, the distortion from Corollary 4.2 is given by

$$d_n(x^n,\hat{x}^n) = -c\cdot\frac{1}{n}\sum_{i=1}^{n}\log_2 P(x_i|\hat{x}_i,x_{i-1}) + d_0(x^n). \tag{4.45}$$

Since $\{\mathbf{X},\hat{\mathbf{X}}\}$ is jointly Gaussian, we have a linear relationship describing $P(X_i|\hat{X}_i,X_{i-1})$

$$X_i = \alpha\hat{X}_i + \beta X_{i-1} + e_i, \quad \forall i \tag{4.46}$$

where $e_i \sim N(0,\sigma_e^2)$ is independent of $\hat{X}_i$ and X_{i-1}. The constants α,β and σ_e^2 have to be determined. Using (4.46) in (4.45), we see that if

$$\begin{aligned} \alpha &= \frac{1}{a},\\ \beta &= -\frac{b}{a},\\ c &= \frac{2\sigma_e^2}{\alpha^2\log_2 e},\\ d_0(x^n) &= -\frac{c}{2}\log_2 2\pi\sigma_e^2, \end{aligned} \tag{4.47}$$

then we obtain

$$d_n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i - (ax_i + bx_{i-1})\right)^2.$$

Hence our guess that a Markov conditional distribution of the form (4.44) achieves the optimum is correct. The optimal rate distortion function is given by

$$\begin{aligned} R_{ff}(D) &= \lim_{N\to\infty}\frac{1}{N}\sum_{x^N,\hat{x}^N} P(x^N,\hat{x}^N)\log_2\prod_{n=1}^{N}\frac{P(x_n|\hat{x}_n,x_{n-1})}{P(x_n|x_{n-1})}\\ &= \sum_{x_1,x_2,\hat{x}_2} P(x_1,x_2,\hat{x}_2)\log_2\frac{P(x_2|x_1,\hat{x}_2)}{P(x_2|x_1)}\\ &= \frac{1}{2}\log_2\frac{\sigma^2(1-\rho^2)}{\sigma_e^2}. \end{aligned} \tag{4.48}$$

σ_e^2 can be determined as follows. We have $\forall i$

$$\begin{aligned} E(X_i^2) &= \sigma^2 \\ E(X_i X_{i-1}) &= \rho\sigma^2 \\ E(X_{i-1}\hat{X}_i) &= (a\rho + b)\sigma^2, \end{aligned} \tag{4.49}$$

where the last equality is obtained by multiplying (4.46) by X_{i-1} and taking expectation. Since the process $\{\mathbf{X}, \hat{\mathbf{X}}\}$ is stationary and ergodic, by the law of large numbers, the distortion

$$\begin{aligned} d_n(x^n, \hat{x}^n) &= \frac{1}{n}\sum_{i=1}^{n} (\hat{x}_i - (ax_i + bx_{i-1}))^2 \\ &\rightarrow E[(\hat{x}_2 - (ax_2 + bx_1))^2] \quad \text{as} \quad n \rightarrow \infty \quad w.p.1. \end{aligned}$$

Therefore the distortion constraint is equivalent to

$$E\left(\hat{X}_2 - (aX_2 + bX_1)\right)^2 \leq D.$$

Setting $E\left(\hat{X}_2 - (aX_2 + bX_1)\right)^2 = D$ and using (4.49), we obtain

$$E(\hat{X}_2^2) = (a^2 + b^2)\sigma^2 + 2ab\rho\sigma^2 - D. \tag{4.50}$$

The final step is to multiply (4.46) by X_i and take expectations to obtain

$$\sigma_e^2 = \frac{D}{a^2},$$

completing the proof. □

We see that the rate-distortion function of the first-order Markov source with distortion criterion $(\hat{X}_n - (aX_n + bX_{n-1}))^2$ depends only on a and not on b. This is reasonable because the decoder knows X_{n-1} while reconstructing $\hat{X}_n$, so X_n is the only 'unknown' quantity.

Also note that when $a = 1$, $b = 0$, the problem becomes to just reconstruct X_n with mean-squared error distortion criterion. Then the rate-distortion function is

$$R_{ff}(D) = \frac{1}{2}\log_2 \frac{(1-\rho^2)\sigma^2}{D}.$$

This can be nicely interpreted as follows. As seen from (4.46), the innovations process of the Gauss-Markov source, N_i is an i.i.d zero-mean Gaussian process with variance $\rho(1-\sigma^2)$. With delay 1 feed-forward, to produce $\hat{X}_n$, the decoder already knows X^{n-1} and the rate R is used merely to convey the innovations process to the decoder. $R_{ff}(D)$ is just the rate-distortion function of the innovations process. This interpretation for the rate-distortion function with feed-forward is true for any source with an i.i.d innovations process as shown in [85].

4.4.2 Channel coding example

In our final example, we consider a simple Markov channel with feedback of the kind described in Section 4.3.1 and evaluate its capacity-cost function. Consider a binary Markov channel with feedback delay 1 characterized by (4.18), i.e., $P^{ch}(Y_i|X^i, Y^{i-1}) = P^{ch}(Y_i|X_i, Y_{i-1})$ for all time instants i. The channel is defined as follows $\forall i$.

$$\begin{aligned} &P^{ch}(Y_i|X_i, Y_{i-1}=1) = \delta_{(Y_i=X_i)}, \\ &P^{ch}(Y_i|X_i, Y_{i-1}=0) = 0.5, \quad X_i, Y_i \in \{0,1\}, \end{aligned} \tag{4.51}$$

In other words, if the channel output at time $i-1$ is 1, we have a noiseless binary channel at time i. If the channel output at time $i-1$ is 0, at time i we have a binary symmetric channel with crossover probability 0.5. The channel is depicted in Figure 4.2.

Suppose we wish to impose a cost function of the following type on the problem.

$$c_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} c(x_i, y_{i-1}) \quad \forall n, \tag{4.52}$$

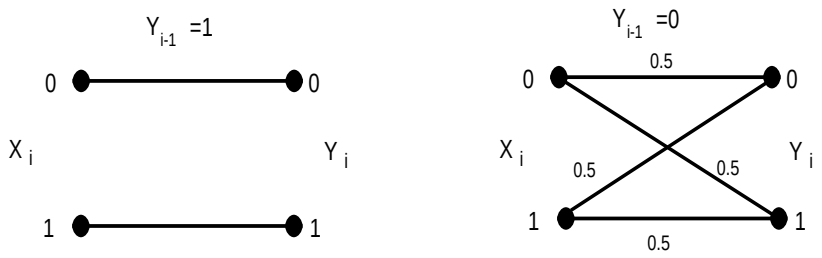

Figure 4.2: Markov channel $\{P(Y_i|X_i, Y_{i-1})\}$. When $Y_{i-1} = 1$, $P(Y_i|X_i, Y_{i-1} = 1)$ is noiseless channel. When $Y_{i-1} = 0$, $P(Y_i|X_i, Y_{i-1} = 0)$ is a BSC with probability $\frac{1}{2}$

where:

1. At time i, if the channel is in the 'good' state ($Y_{i-1} = 1$),

$$c(X_i, Y_{i-1} = 1) = \begin{cases} \alpha_0 & \text{if } x_i = 0 \\ \alpha_1 & \text{if } x_i = 1, \end{cases} \tag{4.53}$$

   for some α_0, α_1.

2. If the channel is in the bad state ($Y_{i-1} = 0$), we impose a constant cost α:

$$c(X_i, Y_{i-1} = 0) = \alpha, \quad X_i \in \{0, 1\}, \tag{4.54}$$

The cost function being defined as above, we would like to evaluate the feedback capacity-cost function $C_{fb}^1(P)$ for a given cost constraint P. One can ask: 'Does a first-order Markov input distribution of the form of (4.20) achieve the optimum in the feedback capacity-cost formula ?' We shall use (4.21) to answer this.

Let the input distribution be given by $\{P_{X_i|Y_{i-1}}\}_{i=1}^{\infty}$ with

$$\begin{aligned} P(X_i = 0|Y_{i-1} = 0) = q_0 \qquad & P(X_i = 0|Y_{i-1} = 1) = q_1 \\ P(X_i = 1|Y_{i-1} = 0) = 1 - q_0 \qquad & P(X_i = 1|Y_{i-1} = 1) = 1 - q_1 \quad \forall i, \end{aligned} \tag{4.55}$$

where q_0, q_1 have to be determined. The joint distribution of the system at any time n is given by

$$P(X^n, Y^n) = \prod_{i=1}^{n} P(X_i, Y_i|Y_{i-1}) = \prod_{i=1}^{n} P(X_i|Y_{i-1}) \cdot P^{ch}(Y_i|X_i, Y_{i-1}), \tag{4.56}$$

with $P(X_i|Y_{i-1})$ and $P^{ch}(Y_i|X_i, Y_{i-1})$ given by and (4.55) and (4.51), respectively. From (4.56), we can determine the marginal $P(Y^n) = \prod_{i=1}^{n} P(Y_i|Y_{i-1})$ as

$$\begin{aligned} &P(Y_i = 0|Y_{i-1} = 0) = 0.5 \qquad P(Y_i = 0|Y_{i-1} = 1) = q_1 \\ &P(Y_i = 1|Y_{i-1} = 0) = 0.5 \qquad P(Y_i = 1|Y_{i-1} = 1) = 1 - q_1, \quad \forall i. \end{aligned} \tag{4.57}$$

Now, using (4.21), we can obtain the conditions for the chosen input distribution to achieve the optimum. Substituting all the possible values for (Y_{i-1}, X_i, Y_i) in (4.21), we see that the conditions are:

$$\begin{aligned} d_0 &= \alpha \\ \lambda \log_2 \frac{1}{q_1} + d_0 &= \alpha_0 \\ \lambda \log_2 \frac{1}{1 - q_1} + d_0 &= \alpha_1 \end{aligned} \tag{4.58}$$

Further the parameter q_1 has to be chosen to satisfy the cost constraint. Since the joint process $\{P_{X^n,Y^n}\}_{n=1}^{\infty}$ (with P_{X^n,Y^n} as in (4.56)) is jointly stationary and ergodic, the cost constraint reduces to

$$E[c(X_i, Y_{i-1})] \leq P.$$

In terms of our joint distribution, this can be written as

(4.59)

$$\begin{aligned} &\sum_{x_i,y_i} P(Y_{i-1} = 0)P(x_i, y_i|Y_{i-1} = 0)c(x_i, 0) + P(Y_{i-1} = 1)P(x_i, y_i|Y_{i-1} = 1)c(x_i, 1) \\ &\stackrel{(a)}{=} P(Y_{i-1} = 0)\alpha + P(Y_{i-1} = 1)[q_1\alpha_0 + (1 - q_1)\alpha_1] \\ &= P, \end{aligned}$$

where we have used (4.51) and (4.55) and the cost function to obtain (a). $[P(Y_{i-1} = 0), \quad P(Y_{i-1} = 1)]$ is just the stationary distribution of the Markov chain $\{Y_i\}$, whose transition probabilities are given by (4.57). This can be computed to be

$$P(Y_{i-1} = 0) = \frac{2q_1}{1 + 2q_1} \qquad P(Y_{i-1} = 1) = \frac{1}{1 + 2q_1}.$$

Using this in (4.59) and rearranging terms, we obtain

$$q_1\alpha_0 + (1-q_1)\alpha_1 + 2q_1\alpha = P(1+2q_1). \tag{4.60}$$

If we fix the cost parameters $\alpha_0, \alpha_1, \alpha$ and the cost constraint P, there are four conditions (given by (4.58) and (4.60)) to be satisfied. We have three variables: λ, d_0, q_1. Hence in general it may not be possible to satisfy all the conditions for predetermined values of $\alpha_0, \alpha_1, \alpha, P$. In other words, one can assert that a first-order Markov input distribution of the form of (4.20) achieves the optimum for this problem only when the system of four equations in three unknowns ((4.58) and (4.60)) has a solution. Of course, if we specify only three among the four parameters $\alpha_0, \alpha_1, \alpha, D$, a solution always exists and the fourth parameter gets automatically determined.

When there exist λ, d_0, q_1 such that (4.58) and (4.60) are satisfied, the feedback capacity-cost function can be evaluated as

$$\begin{aligned}
C_{fb}^1(P) = \underline{I}(X \to Y) &= \lim_{N\to\infty} \frac{1}{N} I(X^N \to Y^N) \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N, y^N} P(x^N, y^N) \log_2 \frac{\prod_{n=1}^N P(y_n|x^n, y^{n-1})}{\prod_{n=1}^N P(y_n|y^{n-1})} \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{x^N, y^N} P(x^N, y^N) \log_2 \prod_{n=1}^N \frac{P(y_n|x_n, y_{n-1})}{P(y_n|y_{n-1})} \\
&= \sum_{y_1, x_1, y_2} P(y_1, x_2, y_2) \log_2 \frac{P(y_2|x_2, y_1)}{P(y_2|y_1)} \\
&= \frac{1}{1+2q_1} \left[q_1 \log_2 \frac{1}{q_1} + (1-q_1) \log_2 \frac{1}{1-q_1} \right] + \frac{2q_1}{1+2q_1} \cdot 0 \\
&= \frac{1}{1+2q_1} h(q_1),
\end{aligned} \tag{4.61}$$

where $h()$ is the binary entropy function. As one might expect, the value of q_0 in the input distribution does not figure in the analysis. This is because when $Y_{i-1} = 0$, both the channel and the cost function behave identically with respect to inputs 0 and 1.

Example 1: Suppose we set $\alpha = 0$, $\alpha_1 = 2\alpha_0$ for any $\alpha_0 > 0$. This implies that when the channel is 'bad', the associated cost is 0; when the channel is 'good', there is a cost > 0 and the cost is higher to keep the channel in the good state. For these parameters, we obtain from (4.58)

$$d_0 = 0, \qquad q_1 = 0.6180.$$

If we set the cost constraint P to satisfy (4.60), we obtain $P = 1.382\alpha_0$ and from (4.61),

$$C_{fb}(1.382\alpha_0) = 0.4291 \text{ bits/channel use.}$$

Example 2: We can also show that the feedback capacity of this channel is a *discontinuous* function of the cost. For any q_1 and $\epsilon > 0$, set

$$\lambda = \epsilon, \quad d_0 = P_0. \tag{4.62}$$

From (4.58), the cost parameters then become

$$\alpha = P_0, \quad \alpha_0 = P_0 + \epsilon \log \frac{1}{q_1}, \quad \alpha_1 = P_0 + \epsilon \log \frac{1}{1-q_1} \tag{4.63}$$

and from (4.60), the cost constraint is equal to $P = P_0 + \epsilon \frac{h(q_1)}{1+2q_1}$. Hence, for the cost parameters in (4.63) the feedback capacity-cost function from (4.61) is

$$C_{fb}\left(P_0 + \epsilon \frac{h(q_1)}{1+2q_1}\right) = \frac{h(q_1)}{1+2q_1}. \tag{4.64}$$

For a fixed q_1, letting $\epsilon \to 0$, the cost parameters in (4.63) all tend to P_0 and the cost constraint $P_0 + \epsilon \frac{h(q_1)}{1-2q_1}$ tends to P_0 as well. However, the right hand side of (4.64) is a function of q_1 alone. We see that the feedback capacity at cost $P_0 + \epsilon \frac{h(q_1)}{1+2q_1}$ varies significantly with q_1 even for arbitrarily small ϵ (cost 'close' to P_0); hence the feedback capacity is a discontinuous function of the cost.

4.5 Conclusion

The rate-distortion function with feed-forward and the channel capacity with feedback involve optimizing the $\limsup_{inprob}$ / $\liminf_{inprob}$ of a multi-letter expression over all valid distributions (*both* stationary and non-stationary). Since the problems studied have memory, these multi-letter optimizations are the only way to characterize the performance limits. Unfortunately, it is not possible to have a simple algorithm (like a Blahut-Arimoto algorithm) to compute the optimization. The rate-distortion expressions involving $\limsup_{inprob}$ rule out simplifying the space of optimization using dynamic programming etc. Hence our approach is especially relevant for problems with feedback/feed-forward.

Our theorems are structural results that relate the structure of the distortion function to the optimal joint distribution. We note that they cannot be used to compute the feedback capacity or rate-distortion function for *all* channels/sources and distortion measures. However, as shown in the examples, we can obtain performance limits for interesting problems which, to the best of our knowledge, are not easy to obtain otherwise.

CHAPTER 5

Multiple Descriptions with Feed-forward

5.1 Introduction

Consider a communication network in which we wish to compress a streaming source of data into packets at one node and transmit them to another node. Assume there is a chance that a packet might be lost and never reach its destination. So we compress each block of data simultaneously into two different packets and send them through different routes. We get a good reconstruction on reception of either packet, but we would like a better reconstruction if both packets are received- in other words, the packets need to refine one another. How should we compress the source into two different descriptions? This, in essence, is the multiple descriptions problem, first posed by Gersho, Ozarow, Witsenhausen and others.

The multiple descriptions set-up is shown in Figure 5.1. $\mathbf{X} = \{X_n\}_{n=1}^{\infty}$ is a source with known distribution. The encoder encodes each block of source samples in two different ways: decoder 1 receives R_1 bits/sample and produces reconstruction $\hat{X}_1$. Similarly, decoder 2 receives R_2 bits/sample and produces $\hat{X}_2$. Decoder 0 receives the full $R_1 + R_2$ bits/sample and produces reconstruction $\hat{X}_0$. Assume suitable distortion measures have been defined for all decoders; let D_1, D_2, D_0 denote the average distortions with which decoders $1, 2$ and 0 are able to reconstruct the source.

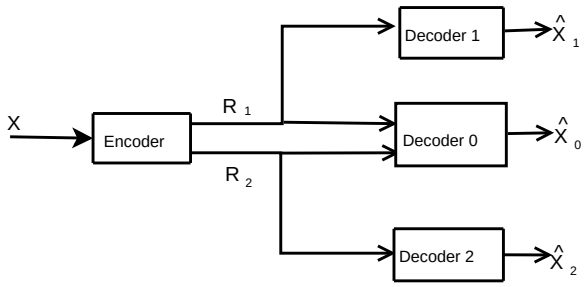

Figure 5.1: The multiple descriptions problem

The problem is to determine the set of all quintuples $(R_1, R_2, D_1, D_2, D_0)$ that are achievable in the usual Shannon sense. This problem has been studied in several notable papers, e.g. [16, 87, 58, 6, 1, 92, 24, 83, 67, 70].

The notion of feed-forward is applicable to multi-terminal source coding problems as well. In Chapter 1, source coding with feed-forward was motivated from a communications perspective as a variant of source coding with side information. The example given was that of compressing a random field to be communicated from one node to another in a network. This field (e.g. an acoustic field) could propagate through the medium at a slow rate and become available at the destination node as side-information with some delay. If the bits from the source were sent to the destination in the form of packets (that could be lost), multiple descriptions coding could be effectively used in this set-up to enhance the quality of reconstruction. Thus it is interesting to study the effect of feed-forward on multiple descriptions source coding. Figure 5.2 shows a multiple descriptions system with feed-forward. Assume switch S_1 is closed and the source samples are sequentially available with a delay k after the indices are sent. To generate $\hat{X}_{1n}$, decoder 1 has knowledge of the index in a codebook (of rate R_1) plus the source samples until time $n-k$. In this paper, we study the achievable quintuples $(R_1, R_2, D_1, D_2, D_0)$ when one or both of S_1 and S_2 are closed.

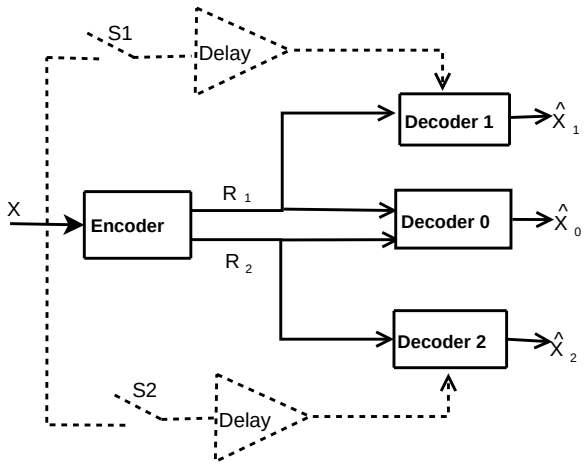

Figure 5.2: The multiple descriptions problem with feed-forward

Source coding with feed-forward is also related closely to prediction. As mentioned in Chapter 1, it was first considered in the context of competitive prediction in [85]. The problem of choosing the best predictor of a random process from an exponentially large class of predictors is equivalent to the source coding problem with feed-forward. The stock market example in Chapter 4 illustrates this connection. The following problem is another example that motivates our study of multiple descriptions with feed-forward. There are four agents Alice, Bob, Carol and Dave. Alice has an equiprobable binary source; Bob, Carol and Dave are interested in reconstructing the source sequence. Bob and Carol each want to reconstruct with the fraction of their errors being at most d, while Dave needs error-free reconstruction. Alice supplies information at rates R_1 and R_2 to Bob and Carol, respectively;, Dave gets the information available to both Bob and Carol. Further assume that after reconstruction of each source sample, Alice reveals to Carol (but not Bob and Dave) the actual value of the sample. The minimum rates of information that Alice would have to supply to Bob and Carol under this scenario is the multiple description rate-distortion region with feed-forward to Carol only. This example is studied in detail

in Section 5.3.

In [65], a simple multiple-description coding scheme was presented for i.i.d. Gaussian sources with feed-forward to *all* decoders (0,1 and 2) with delay $k = 1$. The coding scheme was shown to achieve the optimal rate-distortion region for the i.i.d Gaussian source with feed-forward. In this paper, we present an achievable rate-region for any discrete memoryless source with arbitrary feed-forward delay k, when one or both of S_1 and S_2 in Figure 5.2 are closed.

Before proceeding, we must mention that in this chapter, we consider independent and identically distributed (i.i.d) sources. In point-to-point source coding, feed-forward does not decrease the rate-distortion function of a discrete memoryless source ([85] and Theorem 3). So the simplest sources with feed-forward that are interesting from a rate-distortion perspective are those with memory. In contrast, for multiple descriptions, our results show that the rate-distortion region can be enlarged with feed-forward even for i.i.d sources. This is due to the multi-terminal nature of the problem. A similar situation occurs in channel coding as well. Feedback does not increase the capacity of a discrete memoryless point-to-point channel [74], but it does increase the capacity of many discrete memoryless multi-terminal channels (see for e.g. [26, 17] for multiple access channels and [60, 23] for broadcast channels).

In Section 5.2, we define the problem formally and state the main theorem of this chapter. The proof of the theorem is deferred to Section 5.4. The prediction problem described earlier in this section is discussed in Section 5.3 to illustrate the use of the theorem. We conclude with a short discussion in Section 5.5.

5.2 Problem statement and main results

Consider a discrete memoryless source X with finite alphabet $\mathcal{X}$. We assume that the source samples $X_n,\ n = 1, 2, \ldots$ are independent and identically distributed (i.i.d) according to a probability mass function $P_X(x)$. Let $\widehat{\mathcal{X}}_0, \widehat{\mathcal{X}}_1, \widehat{\mathcal{X}}_2$ denote the finite reconstruction spaces of decoder 0,1 and 2, respectively. Each reconstruction has an associated single-letter distortion measure $d_m : \mathcal{X} \times \widehat{\mathcal{X}}_m \to \mathbb{R}, \quad m = 0, 1, 2.$ The distortion on N-length sequences is the average of the per-letter distortions. For all $x^N \in \mathcal{X}^N, \hat{x}^N \in \widehat{\mathcal{X}}^N$,

$$d_m(x^N, \hat{x}_m^N) \triangleq \frac{1}{N} \sum_{n=1}^{N} d_m(x_n, \hat{x}_{mn}), \quad m = 0, 1, 2.$$

5.2.1 Feed-forward to only one decoder

Without loss of generality assume S_1 is open and S_2 is closed in Figure 5.2.

Definition 5.1. An $(N, 2^{NR_1}, 2^{NR_2})$ multiple description code of block length N and rates (R_1, R_2), with delay k feed-forward to decoder 2, consists of:

1. Encoder mappings $e_m : \mathcal{X}^N \to \{1, \ldots, 2^{NR_m}\}, \quad m = 1, 2.$

2. Mappings for decoders 0 and 1:

$$g_0 : \{1, \ldots, 2^{NR_1}\} \times \{1, \ldots, 2^{NR_2}\} \to \widehat{\mathcal{X}}_0^N$$

$$g_1 : \{1, \ldots, 2^{NR_1}\} \to \widehat{\mathcal{X}}_1^N$$

3. A sequence of mappings for decoder 2:[1]

$$g_{2n} : \{1, \ldots, 2^{NR_2}\} \times \mathcal{X}^{n-k} \to \widehat{\mathcal{X}}_2, \quad n = 1, \ldots, N.$$

[1]It is understood that for $n \leq k$, $\mathcal{X}^{n-k}$ is the empty set.

The encoder maps each N-length source sequence to a pair of indices in $\{1,\ldots,2^{NR_1}\}\times\{1,\ldots,2^{NR_2}\}$. The decoders receive their respective indices. In addition, decoder 2 has access to the source samples until time $(n-k)$ to reconstruct the nth sample,. Achievable rates are defined in the usual Shannon sense.

Definition 5.2. (R_1,R_2) is an achievable rate pair for distortion (D_0,D_1,D_2) if there exists a sequence, indexed by N, of $(N,2^{NR_1},2^{NR_2})$ multiple description codes with feed-forward delay k, such that for sufficiently large N,

$$Ed_m(X^N,\hat{X}_m^N)\le D_m,\quad m=0,1,2.$$

The rate distortion region $R(D_0,D_1,D_2)$ is the closure of the set of achievable rate pairs for distortion (D_0,D_1,D_2).

Our main result is the following theorem.

Theorem 13. *A quintuple (R_1,R_2,D_0,D_1,D_2) is achievable - with delay k feed-forward to decoder 2 only- if there exist random variables $U,\hat{X}_1,\hat{X}_2,\hat{X}_0$ jointly distributed with the source X such that*

$$\begin{aligned}
&R_1 > I(X;\hat{X}_1U)\\
&R_2 > I(X;\hat{X}_2|U) + \ max\{0,R_1-I(X\hat{X}_2;\hat{X}_1|U)\}\\
&R_1+R_2 > I(X;\hat{X}_1U)+I(X;\hat{X}_2|U)+I(X;\hat{X}_0|\hat{X}_1\hat{X}_2U)\\
&\qquad + I(\hat{X}_1;\hat{X}_2|XU) + \ max\ \{0,R_1-I(X\hat{X}_2;\hat{X}_1|U)\}\\
&Ed_m(X;\hat{X}_m)\le D_m,\quad m=0,1,2
\end{aligned}$$

The proof of the theorem is given in Section 5.4. Notice that the rate-region specified by the theorem does not depend on the feed-forward delay k, i.e., the region is achievable for any finite delay k. We can compare this rate region with the rates achievable for multiple descriptions without feed-forward. The multiple descriptions

rate-distortion region (without feed-forward) is known only for certain special cases (see [58, 16, 1, 24]). The best known achievable region for the general two-channel multiple descriptions problem for an i.i.d source is due to Zhang and Berger [92]. We reproduce this rate-region below in a slightly modified, but equivalent, form.

Theorem 14 ([92]). *A quintuple $(R_1, R_2, D_0, D_1, D_2)$ is achievable (without feed-forward) if there exist random variables $U, \hat{X}_1, \hat{X}_2, \hat{X}_0$ jointly distributed with the source X such that*

$$R_1 > I(X; \hat{X}_1 U), \qquad R_2 > I(X; \hat{X}_2 U)$$

$$R_1 + R_2 > I(X; \hat{X}_1 U) + I(X; \hat{X}_2 U) + I(X; \hat{X}_0 | \hat{X}_1 \hat{X}_2 U) + I(\hat{X}_1; \hat{X}_2 | XU)$$

$$E d_m(X; \hat{X}_m) \leq D_m, \quad m = 0, 1, 2$$

To see that Theorem 13 enlarges the no-feed-forward rate region of Theorem 14, consider any set of random variables $U, \hat{X}_1, \hat{X}_2, \hat{X}_0$ jointly distributed with X. Set $R_1 = I(X; \hat{X}_1 U) + \epsilon$ for some small $\epsilon > 0$. We can have one of two situations:

(a) $R_1 = I(X; \hat{X}_1 U) + \epsilon \leq I(X\hat{X}_2; \hat{X}_1 | U)$: In this case, from Theorem 13,

$$R_2 = I(X\hat{X}_1; \hat{X}_2 | U) + I(X; \hat{X}_0 | \hat{X}_1 \hat{X}_2 U) + \epsilon$$

is achievable. This represents a savings of $I(U; X)$ bits/sample over the minimum R_2 without feed-forward (specified by Theorem 14).

(b) $R_1 = I(X; \hat{X}_1 U) + \epsilon > I(X\hat{X}_2; \hat{X}_1 | U)$: We now have

$$R_2 = I(X\hat{X}_1; \hat{X}_2 | U) + I(X; \hat{X}_0 | \hat{X}_1 \hat{X}_2 U) + [I(X; U) - I(\hat{X}_2; \hat{X}_1 | XU)] + \epsilon$$

is achievable, a savings of $I(\hat{X}_2; \hat{X}_1 | XU)$ bits/sample over the no-feed-forward case. Of course, the potential savings in rate may be greater since we have only presented an achievable rate region.

5.2.2 Feed-forward to both decoders 1 and 2

Switches S_1 and S_2 in Figure 5.2 are *both* closed. An $(N, 2^{NR_1}, 2^{NR_2})$ multiple description code with delay k feed-forward is defined in the same way as the previous subsection, except that now both decoder 1 and 2 are defined by a sequence of mappings. In addition to the index, both decoders 1 and 2 have access to the source samples until time $(n-k)$.

Achievable rates are defined as before. Clearly, the region of Theorem 13 is achievable. The rate region obtained by switching the roles of R_1 and R_2 in Theorem 13 is also achievable. Thus the convex hull of the union of these two regions is a (possibly larger) achievable rate-region.

5.3 Example

Consider an i.i.d binary source X with pmf $P_X(0) = P_X(1) = 1/2$. The reconstruction spaces are all binary and the distortion measures are Hamming, i.e., $d(x, \hat{x}_m) = \delta_{x \neq \hat{x}_m}, \quad m = 0, 1, 2$. Suppose decoders 1 and 2 want to reconstruct X with distortion d, while decoder 0 needs error-free reconstruction. We want to characterize the minimum sum-rate

$$r_{sum}(d) \triangleq \inf\{R_1 + R_2 : (R_1, R_2, 0, d, d) \text{ achievable}\}. \tag{5.1}$$

A lower bound to $r_{sum}(d)$ without feed-forward was obtained in [92, Theorem 3, Section VIII] [2]:

$$r_{sum}(d)_{no-ff} \geq 2 - h\left(\frac{4d + 1 - \sqrt{12d^2 - 4d + 1}}{2}\right). \tag{5.2}$$

[2] There appears to be a typo in the statement of the result in [92, Theorem 3]. The correct version (given here) can be obtained from the proof of that theorem.

Let us now assume only decoder 2 gets feed-forward with delay k. Let U be a binary-valued random variable and fix the conditional distribution $P_{U,\hat{X}_1,\hat{X}_2,\hat{X}_0|X} = P_{U|X}P_{\hat{X}_1,\hat{X}_2|XU}P_{\hat{X}_0|XU\hat{X}_1\hat{X}_2}$ as follows.

Fix a parameter $D_0, \quad 0 \leq D_0 \leq 1$ and define

$$\begin{aligned} P_{U|X}(0|0) &= P_{U|X}(1|1) = 1 - D_0 \\ P_{U|X}(0|1) &= P_{U|X}(1|0) = D_0. \end{aligned} \tag{5.3}$$

$P_{\hat{X}_1,\hat{X}_2|XU}$ is defined as

$$\begin{aligned} P_{\hat{X}_1,\hat{X}_2|XU}(00|00) &= P_{\hat{X}_1,\hat{X}_2|XU}(11|11) = 1 \\ P_{\hat{X}_1,\hat{X}_2|XU}(01|01) &= P_{\hat{X}_1,\hat{X}_2|XU}(10|01) = d/D_0 \\ P_{\hat{X}_1,\hat{X}_2|XU}(00|01) &= 1 - 2d/D_0 \\ P_{\hat{X}_1,\hat{X}_2|XU}(01|10) &= P_{\hat{X}_1,\hat{X}_2|XU}(10|10) = d/D_0 \\ P_{\hat{X}_1,\hat{X}_2|XU}(11|10) &= 1 - 2d/D_0 \end{aligned} \tag{5.4}$$

$\hat{X}_0$ is a function of $(U, \hat{X}_1, \hat{X}_2)$:

$$\hat{X}_0 = \begin{cases} \hat{X}_1 & \text{if } (\hat{X}_1 = \hat{X}_2) \\ 1 - U & \text{if } (\hat{X}_1 \neq \hat{X}_2) \end{cases} \tag{5.5}$$

It is easy to check that this joint distribution achieves the distortion triple ($D_1 = d, D_2 = d, D_0 = 0$). Using this in Theorem 13, we can obtain an achievable rate-region when only decoder 2 receives feed-forward. The relevant information quanti-

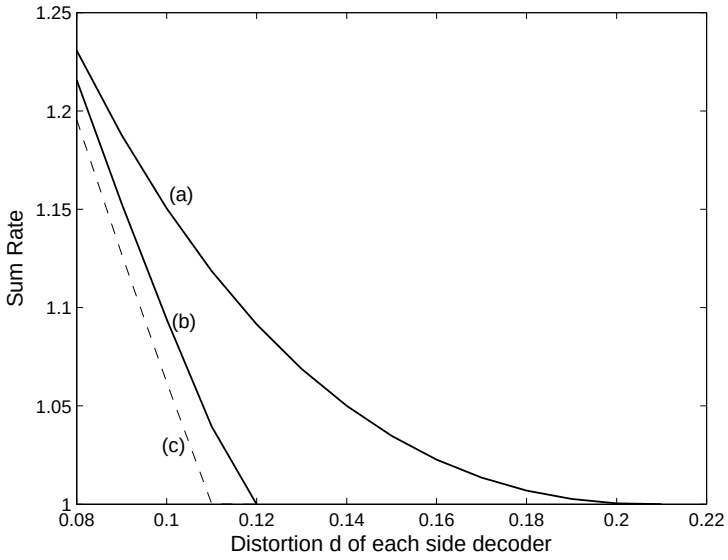

Figure 5.3: (a) Zhang-Berger lower bound on $r_{sum}(d)$ without FF, (b) Achievable sum-rate with FF to one decoder , (c) Rate-distortion lower bound on $r_{sum}(d)$ with FF.

ties are calculated below, with $h(.)$ used to denote the binary entropy function.

$$
\begin{aligned}
& I(X;U) = H(U) - H(U|X) = 1 - h(D_0). \\
& I(\hat{X}_2;\hat{X}_1|XU) = H(\hat{X}_1|XU) - H(\hat{X}_1|\hat{X}_2XU) \\
& \qquad\qquad = D_0 h(\frac{d}{D_0}) - D_0\left(1-\frac{d}{D_0}\right)h(\frac{d}{D_0-d}). \\
& I(X;\hat{X}_2|U) = I(X;\hat{X}_1|U) = H(\hat{X}_1|U) - H(\hat{X}_1|UX) \\
& \qquad\qquad = h(D_0-d) - D_0 h(\frac{d}{D_0}). \\
& I(X;\hat{X}_0|\hat{X}_1\hat{X}_2U) = 0.
\end{aligned}
\tag{5.6}
$$

(5.6) contains all the expressions required to compute the rate-region of Theorem 13. Thus for each d, we can select the value D_0 to yield the best rate-constraint and obtain an achievable upper bound to $r_{sum}(d)$ in (5.1) (with feed-forward to only one decoder). This is plotted in graph (b) of Figure 5.3 for distortions $d \geq 0.08$. Graph (a) is the Zhang-Berger lower bound (5.2) to $r_{sum}(d)$ without feed-forward. We see that for all the distortions considered, feed-forward to one decoder yields achievable rates smaller than the optimal no feed-forward rate. Since decoders 1 and

2 produce reconstructions with distortion d, R_1 and R_2 have to each be greater than the Shannon rate-distortion function $R(d) = 1 - h(d)$. This is true both with and without feed-forward. Thus a simple lower bound to $r_{sum}(d)$ with feed-forward is $r_{sum}(d) > 2(1 - h(d))$, which is plotted in graph (c) of Figure 5.3.

Of particular interest is the situation when the sum rate $R_1 + R_2 = 1$. This is the case of *no excess rate* to the central decoder [1]. Setting $D_0 = 0.25945$, we see from Theorem 1 that $(R_1 = 0.5, R_2 = 0.5)$ can achieve $d = 0.12$ with feed-forward to one decoder. In comparison, it was shown in [6] that with rates of (0.5,0.5) and no feed-forward, the minimum achievable distortion at each side-decoder is $(\sqrt{2} - 1)/2 = 0.207$.

5.4 Proof of Theorem

Assume delay k feed-forward, i.e. each source sample is available at the decoder $N + k$ time units after it is available to the encoder (N will be a measure of block-length in the coding scheme). First fix the joint distribution

$$P_X(x) \cdot P_{U,\hat{X}_1,\hat{X}_2,\hat{X}_0|X}(u, \hat{x}_1, \hat{x}_2, \hat{x}_0|x).$$

In the sequel, upper-case letters will be used for random variables and lower-case letters for their realizations. Vectors will be denoted in bold letters.

To prove the theorem, we shall use the properties of strongly ϵ-typical sequences [19]. Length-$N/2$ vectors $x^{N/2}, \hat{x}_1^{N/2}, \hat{x}_2^{N/2}$ are said to be jointly typical if their joint type (composition) is approximately $P_{X,\hat{X}_1,\hat{X}_2}$. The set of all jointly ϵ-typical tuples $(X^{N/2}, \hat{X}_1^{N/2}, \hat{X}_2^{N/2})$ is denoted $T_\epsilon(\mathbf{X}, \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2)$. Similar definitions of typicality hold for other joint and conditional distributions.

We divide the source sequence into a large number of blocks, say B blocks, with each block containing $\frac{N}{2}$ source symbols. To exploit the feed-forward, we shall use

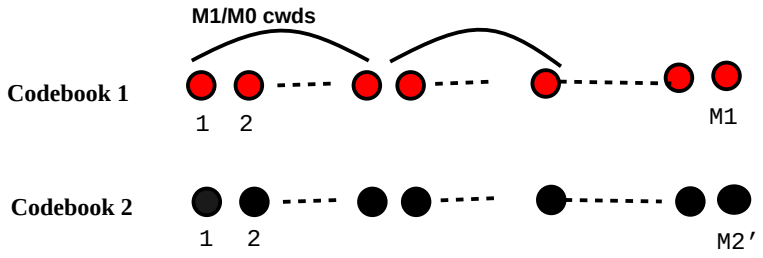

Figure 5.4: Codebook cells for decoder 1

a block-Markov superposition strategy [17, 86] covering two adjacent blocks. The ideas of non-random binning and restricted encoding, introduced in [86], will be used in the proof.

Random Coding: Let $M_0 = 2^{\frac{N}{2}R_0}$, $M_1 = 2^{\frac{N}{2}R_1}$ and $M_2^{'} = 2^{\frac{N}{2}R_2^{'}}$. Choose $\mathbf{U}(1), \ldots, \mathbf{U}(M_0)$ independently according to a uniform distribution over the set $T_\epsilon(\mathbf{U})$ of all the ϵ-typical $N/2$-vectors $\mathbf{U}$. For each $\mathbf{U}(i)$, choose a codebook of length-$N/2$vectors $\hat{\mathbf{X}}_1^i(1), \ldots, \hat{\mathbf{X}}_1^i(M_1)$, independently according to a uniform distribution over the set $T_\epsilon(\hat{\mathbf{X}}_1|\mathbf{U}(i))$. Similarly choose $\hat{\mathbf{X}}_2^i(1), \ldots, \hat{\mathbf{X}}_2^i(M_2^{'})$ from $T_\epsilon(\hat{\mathbf{X}}_2|\mathbf{U}(i))$.

$\mathbf{U}$ can be thought of as a 'cloud center' conditioned on which reconstructions are produced at decoders 1 and 2. The coding strategy uses the feed-forward to decoder 2 to convey $\mathbf{u}$ 'cheaply' to the decoders. To facilitate this, we partition each $\hat{\mathbf{X}}_1^i$ codebook into M_0 disjoint cells, so that each cell has M_1/M_0 elements. The codebook structures are shown in Figure 5.4. We have assumed for simplicity that M_1/M_0 is an integer.

Encoding: We encode a source sequence $\mathbf{x}$ spanning B blocks, each block containing $N/2$ source symbols. We denote the bth block by $\mathbf{x}_b$, $b = 1, \ldots B$. Thus

$$\mathbf{x} = [x_1, x_2, \ldots, x_{BN/2}] = [\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_B].$$

Step 0: Set $\mathbf{u}_1 = \mathbf{U}(1)$.

Step b ($b = 1, \ldots, B$): Assuming $\mathbf{u}_b$ is known to be equal to $\mathbf{U}(i)$, encode $\mathbf{x}_b$ as

Table 5.1: Time-line of events at encoder and decoder with feed-forward with $k = 1$

Time instant	1	...	$N/2$	...	$2N/2$	...	$N(b+1)/2$	·
Source	X_1	...	$X_{N/2}$	...	$X_{2N/2}$	...	$X_{N(b+1)/2}$	·
Encoder					(w_{11}, w_{21})		(w_{1b}, w_{2b})	
FF at decoder 2	-	-	-	-	-	$X_1X_2\ldots$	...	...
Reconstruction					$\hat{\mathbf{x}}_{11}, \hat{\mathbf{x}}_{21}$		$\hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b}$	

follows. Observe the next length-$N/2$ block $\mathbf{x}_{b+1}$ and find a $j \in \{1, \ldots, M_0\}$ such that $(\mathbf{x}_{b+1}, \mathbf{U}(j)) \in T_\epsilon(\mathbf{X}, \mathbf{U})$. Set $\mathbf{u}_{b+1} = \mathbf{U}(j)$. If no such j is found or if $b = B$, set $\mathbf{u}_{b+1} = \mathbf{U}(1)$. So we have $\mathbf{u}_b = \mathbf{U}(i), \mathbf{u}_{b+1} = \mathbf{U}(j)$.

Pick $(w_{1b}, w'_{2b}) \in \{1, \ldots, M_1\} \times \{1, \ldots, M_2'\}$ such that $(\mathbf{x}_b, \mathbf{u}_b, \hat{\mathbf{X}}_1^i(w_{1b}), \hat{\mathbf{X}}_2^i(w'_{2b})) \in T_\epsilon(\mathbf{X}, \mathbf{U}, \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2)$ *and* $\hat{\mathbf{X}}_1^i(w_{1b})$ belongs to the jth cell of the $\hat{\mathbf{X}}_1^i$ codebook. If no such (w_{1b}, w'_{2b}) is found, set w_{1b} to a random index in the jth cell of the $\hat{\mathbf{X}}_1^i$ codebook, and similarly set w'_{2b} to a random index in the $\hat{\mathbf{X}}_2^i$ codebook.

Note that we restrict ourselves to one cell within the $\hat{\mathbf{X}}_1^i$ codebook. Restricted encoding enables decoder 2 to take advantage of the feed-forward. Decoders 1 and 2 will receive w_{1b} and w'_{2b}, respectively and produce reconstructions $\hat{\mathbf{x}}_{1b}$ and $\hat{\mathbf{x}}_{2b}$. Later, decoder 2 learns $\mathbf{x}_b$ precisely through feed-forward and tries to decode $\hat{\mathbf{x}}_{1b}$ using $(\mathbf{x}_b, \hat{\mathbf{x}}_{2b})$. To facilitate this, the encoder might need to send some extra bits to decoder 2 (in addition to w'_{2b}). These extra bits sent to decoder 2 are represented as an additional index w''_{2b} from an appropriate codebook of size $2^{R_2'' N/2}$. The total rate R_2 sent to decoder 2 is thus $R_2' + R_2''$. In summary, the encoder sends w_{1b} to decoder 1 and (w'_{2b}, w''_{2b}) to decoder 2.

Decoding: Since there is a growing amount of information available at decoder 2(due to feed-forward), the time-line of observations at the encoder and decoder is important. Recall that a source sample is available to the decoder $N + k$ time units after it is produced. The time-line of various events at the encoder and decoder with

$k = 1$ is shown in Table 5.1.

Step b ($b = 1, \ldots, B$): To produce the indices corresponding to block b, the encoder uses blocks $\mathbf{x}_b$ and $\mathbf{x}_{b+1}$. At time $(b+1)N/2$, the source has produced $b+1$ blocks, indices $w_{1b}, (w'_{2b}, w''_{2b}), (w_{1b}, w'_{2b}, w''_{2b})$ are received by decoders $1, 2, 0$, respectively. As will be explained, $\mathbf{u}_b = \mathbf{U}(i)$ has been decoded by all decoders just before time $(b+1)N/2$. The appropriate codebooks $\hat{\mathbf{X}}_1^i, \hat{\mathbf{X}}_2^i$ are used and reconstructions $\hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b}$ are produced using w_{1b}, w'_{2b}, respectively. The generation of $\hat{\mathbf{x}}_{0b}$ is described at the end of the proof.

By time instant $(b+2)N/2$, decoder 2 has received the first b blocks of source samples $\mathbf{x}_1, \ldots, \mathbf{x}_b$ through feed-forward (each block has length $N/2$; we can assume $N >> k$, so that receiving $N/2 - k$ source samples is equivalent to receiving the entire block). Decoder 2 then tries to find $\hat{\mathbf{x}}_{1b}$ from the $\hat{\mathbf{X}}_1^i$ codebook such that $(\mathbf{x}_b, \mathbf{u}_b, \hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b}) \in T_\epsilon(\mathbf{X}, \mathbf{U}, \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2)$. If there is more than one $\hat{\mathbf{x}}_{\mathbf{1b}}$ satisfying the condition, w''_{2b} resolves the list. The cell number j^* of $\hat{\mathbf{x}}_{1b}$ determines $\mathbf{u}_{b+1} = \mathbf{U}(j^*)$. Thus by time instant $(b+2)N/2$, all three decoders know $\mathbf{u}_{b+1}$.

Probability of Error: For our coding strategy, we will declare an error in block b ($b = 1, \ldots, B$) if one or more of the following events occur.

1. Event E_1: The source vector $\mathbf{x}_b$ is not a typical sequence with respect to P_X.
2. E_2: The encoder cannot find $j \in \{1, \ldots, M_0\}$ such that $\mathbf{U}(j)$ is jointly typical with $\mathbf{x}_{b+1}$.
3. E_3: Assuming $\mathbf{u}_b = \mathbf{U}(i), \mathbf{u}_{b+1} = \mathbf{U}(j)$, the encoder cannot find a $(\hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b})$ such that $(\mathbf{x}, \hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b}, \mathbf{u}_b)$ is jointly typical *and* $\hat{\mathbf{x}}_{1b}$ is in the jth cell of its codebook.
4. E_4: Decoder 2 is unable to decode $\hat{\mathbf{x}}_{1b}$ correctly with knowledge of $(\mathbf{x}_b, \hat{\mathbf{x}}_{2b})$ and w''_{2b}.

We bound the probability of each each event for sufficiently large N as follows.

Consider any $\epsilon > 0$. With high probability $\mathbf{x}_b$ is typical with respect to P_X. Thus $P(E_1) < \epsilon/4$.

For $b = 1, \ldots, B-1$,there exists a codebook $\{\mathbf{U}(j),\ j \in \{1, \ldots, M_0\}\}$ such that with high probability, at least one codeword is jointly typical with $\mathbf{x}_{b+1}$ iff $M_0 > 2^{I(X;U)N/2}$. Hence $P(E_2) < \epsilon/4$ if

$$R_0 > I(X;U). \tag{5.7}$$

To compute $P(E_3)$, we first note that given $\mathbf{u}_b = \mathbf{U}(i), \mathbf{u}_{b+1} = \mathbf{U}(j)$, we need to find an $\hat{\mathbf{x}}_{1b}$ from the jth cell of $\hat{\mathbf{X}}_1^i$ codebook (a cell has $2^{(R_1-R_0)N/2}$ codewords) and an $\hat{\mathbf{x}}_{2b}$ from the $\hat{\mathbf{X}}_2^i$ codebook ($2^{R_2^{'}N/2}$ codewords) such that $(\hat{\mathbf{x}}_{1b}, \hat{\mathbf{x}}_{2b}) \in T_\epsilon(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2|\mathbf{X}, \mathbf{U})$. Using arguments similar to the proof in [16], we can show that this is possible with high probability (i.e., $P(E_3) < \epsilon/4$) if

$$\begin{aligned} &R_1 - R_0 > I(X;\hat{X}_1|U) \qquad R_2^{'} > I(X;\hat{X}_2|U) \\ &R_1 - R_0 + R_2^{'} > I(X;\hat{X}_1|U) + I(X;\hat{X}_2|U) + I(\hat{X}_1;\hat{X}_2|XU) \end{aligned} \tag{5.8}$$

Assuming there was no encoding error, i.e. $(E_1 \cup E_2 \cup E_3)^c$, the $\hat{\mathbf{X}}_{1b}$ chosen by the encoder is jointly typical with $(\mathbf{x}_b, \hat{\mathbf{x}}_{2b})$. The probability that another random $\hat{\mathbf{X}}_{1b} \in T_\epsilon(\hat{\mathbf{X}}_1|\mathbf{U})$ is jointly typical with a random pair $(\mathbf{X}_b, \hat{\mathbf{X}}_{2b}) \in T_\epsilon(\mathbf{X}, \hat{\mathbf{X}}_2|\mathbf{U})$ is approximately $2^{-I(\hat{X}_1;X\hat{X}_2|U)N/2}$ for large $N/2$. Thus the number of other $\hat{\mathbf{X}}_1$ codewords that are jointly typical with the known pair $(\mathbf{x}_b, \hat{\mathbf{x}}_{2b})$ is approximately

$$M_1 \cdot 2^{-I(\hat{X}_1;X\hat{X}_2|U)N/2} = 2^{(R_1 - I(\hat{X}_1;X\hat{X}_2|U))N/2} \tag{5.9}$$

Thus if $R_1 > I(\hat{X}_1; X\hat{X}_2|U)$, w''_{2b} has to resolve a list whose size is given by (5.9).Hence we can have $P(E_4) < \epsilon/4$ if the rate $R_2^{''}$ of the extra index satisfies

$$R_2^{''} > \max\{0, R_1 - I(\hat{X}_1; X\hat{X}_2|U)\} \tag{5.10}$$

Assume (5.7), (5.8) and (5.10) are satisfied. From the arguments above and the union bound, we see that P_{be}, the probability of error in block b, satisfies $P_{be} < \epsilon, \quad b = 2, \ldots, B$. It should be noted here that in the first step, we arbitrarily set $\mathbf{u}_1 = \mathbf{U}(1)$. In general, $\mathbf{u}_1$ will not be jointly typical with $\mathbf{x}_1$. Consequently, for the first block alone $P(E_2) = 1$. The average probability of error over B blocks is

$$P_e = \frac{1}{B}\sum_{b=1}^{B} P_{eb} < \frac{1}{B}\left(1 + \epsilon \ldots + \epsilon\right) = \frac{1}{B} + \frac{(B-1)\epsilon}{B} < 2\epsilon$$

for sufficiently large B.

Finally, conditioned on the knowledge of central decoder 0, $\mathbf{X}$ can be quantized to $\hat{\mathbf{X}}_0$. It can be shown[16] that the extra rate needed by the central decoder is $I(X; \hat{X}_0 | \hat{X}_1, \hat{X}_2, U)$. This overhead needs to be shared between the rates R_1 and R_2. We combine this shared overhead with the rates specified by (5.7), (5.8) and (5.10), and recognize that $R_2 = R_2^{'} + R_2^{''}$ to obtain the rate region of Theorem 13.

5.5 Conclusion

In this chapter, the two-channel multiple descriptions problem for an i.i.d source, with feed-forward to one or both side-decoders was considered. A single-letter achievable rate-region was derived using a block-Markov superposition source coding strategy. This achievable region is larger than the best known rate-region for multiple descriptions without feed-forward. In point-to-point source coding, feed-forward does not decrease the rate-distortion function of an i.i.d source. In contrast, we presented an example which showed that the derived region can be larger than the optimal multiple description rate region without feed-forward.

We note that our coding scheme used to prove the result uses feed-forward to one decoder only. It is worth exploring how we can do better if we have feed-forward to all *three* decoders (including the central one). This is especially interesting because

it has been shown [65] that for an i.i.d Gaussian source with feed-forward to all three decoders, we do not need any excess rate, i.e., we can achieve the optimal rate-distortion functions at all three decoders simultaneously.

CHAPTER 6

Summary and Future Directions

6.1 Summary

In this thesis, two related problems- source coding with feed-forward and channel coding with feedback- were studied from an information-theoretic perspective. Feed-forward in source coding introduces a dynamic aspect to the decoder, while feedback in channels introduces dynamics in the encoder.

In Chapter 2, a source coding problem with noiseless feed-forward was formally defined. The optimal rate-distortion function and error exponent of a general source with feed-forward were characterized using directed information, a variant of the more well-known mutual information. An interpretation of directed information was given, which explained why it is the right quantity to characterize the feed-forward rate-distortion function. This interpretation helped generalize the notion of directed information. Using this, we obtained the rate-distortion function for source coding with arbitrarily delayed feed-forward.

A similar interpretation of directed information in the context of channel coding with feedback was presented in Chapter 3. Using this interpretation, we obtained the feedback capacity of a channel with state, with the state information available at both transmitter and receiver (possibly with delay). We also established a source-channel

separation theorem for general sources and channels with feedback and feed-forward, under suitable constraints on the source and channel distributions.

Chapter 4 addressed the problem of computing the optimum rate-distortion function and capacity expressions. These involve optimizing directed information over an infinite dimensional space of distributions. It is difficult to optimize these 'multi-letter' expressions in general. We presented a different approach to this optimization problem and obtained the structure of the distortion (cost, resp.) function which achieves the optimum for a given joint distribution on the source (channel, resp.) coding system. Several examples were provided to illustrate the utility of this approach.

Finally, in Chapter 5, we explored the role of feed-forward in multiple-description source coding. In contrast to point-to-point source coding, we found that feed-forward can strictly improve the multiple descriptions rate-distortion region even for an i.i.d source. We obtained a single-letter achievable rate-region for multiple descriptions of an i.i.d source with feed-forward.

6.2 Future Directions

Several interesting extensions and problems related to feedback and feed-forward suggest themselves. Some were mentioned at the end of the individual chapters. A few others are discussed below.

- **Noisy feedback/feed-forward**. In a channel with noiseless feedback, the transmitter knows (with some delay) the exact output observed by the receiver. In practice, the feedback available at the transmitter is a usually a noisy version of the channel output. There has been work on obtaining error exponents for discrete memoryless channels with noisy feedback (see, for example, [43, 22]).

However, there is no characterization of achievable rates for a general channel with noisy feedback. In noiseless feedback, the encoder and decoder are synchronized in the sense that the encoder has complete knowledge of everything the decoder observes. This is not the case with noisy feedback which makes it a significantly harder problem. Just like directed information is the key to understanding noiseless feedback, new insights and definitions of information measures might be needed to deal with noisy feedback.

- **Multiple access channels with feedback**. In a two-user multiple access channel (MAC), there are two transmitters that use the channel to transmit data to a single receiver. This situation occurs, for example, in the uplink of the cellular system. Kramer [47] characterized the feedback capacity region of this channel in terms of directed information. However, this rate region involves multi-letter expressions (even for a discrete memoryless MAC) that are difficult to compute. On the other hand, a single-letter achievable rate region for a feedback MAC was obtained by Cover and Leung [17]. This region is easy to compute, but is known to be strictly smaller than the capacity region. A natural question arises- can we improve the Cover-Leung region using two-letter and three-letter characterizations (as opposed to single-letter), and maybe even approach the multi-letter capacity expression by increasing the dimension? A careful examination of the Cover-Leung coding scheme yields some preliminary ideas on how it could be improved. The scheme essentially transmits a pair of the two users' messages using two successive blocks of transmission. The message rates of the two users are too high for the receiver to decode the messages after the first block of transmissions, but are low enough to allow the encoders to learn the messages of one another using the feedback. The encoders

then *cooperate* in the next block of transmissions to resolve the decoder's uncertainty. One can now ask- what if the users transmit at rates that are too high to decode each other's message with the feedback? In this case, each user has only *partial* information about the message of the other user. The degree of cooperation is reduced, but they can still send correlated information to resolve the decoder's uncertainty. The methods of [14] and [68] could be useful in exploring if this idea yields higher achievable rates than the Cover-Leung region. We must mention here that in [9] the Cover-Leung rate region is extended using a different approach.

- **Broadcast channels with feedback**. A broadcast channel is a model with a single input and two outputs (users). There is a separate message that each user wishes to transmit, and possibly a common message as well. The broadcast channel is an important model that forms a part of many communication networks (e.g. the down-link in cellular networks). The problem of broadcast channels with feedback remains largely unexplored. It is known that feedback can enhance the capacity region of memoryless broadcast channels [23, 60], but no characterization of achievable rates for a discrete, memoryless braodcast channel with feedback exists. Obtaining a singe-letter achievable rate-region for broadcasting with feedback is an important open problem. This seems to be significantly harder than the feedback MAC. In a MAC, one can see that feedback enables *cooperation* between the transmitters which results in improved rates. It is not clear how feedback helps the two receivers 'cooperate' in a broadcast channel. One possibility is that feedback might enable the transmitter to send some common information to resolve the two users' uncertainty.

- **Other applications of directed information**. It seems likely that directed

information could be useful in problems very different from communications with feed-forward and feedback. As mentioned earlier, quantities that closely resemble directed information have been defined to understand the casual relationship between time-series (see, for example, [30] that attempts to characterize causal dependence between economic time-series). It will be interesting to explore other applications where directed information or similarly defined quantities might help capture the direction of information flow.

APPENDICES

APPENDIX A

Proofs for Chapter 2

A.1 Proof of Lemma 2.4(AEP)

The proof is similar to that of the Shannon-McMillan Breiman theorem in [15, 2]. We first state the definitions and three lemmas required for the proof. Recall that

$$\vec{P}(\hat{x}^N|x^N) = \prod_{i=1}^{N} P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}),$$
$$\vec{P}(x^N|\hat{x}^N) = \prod_{i=1}^{N} P(x_i|\hat{x}^i, x^{i-1}).$$

We want to show that

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) \to \vec{H}(\hat{X}||X) \triangleq \lim_{N\to\infty} H(\hat{X}_N|X^{N-1}, \hat{X}^{N-1}) \tag{A.1}$$

Definition A.1. Let

$$\vec{H}^{\infty}(\hat{X}||X) = E\left[-\log P(\hat{X}_0|\hat{X}_{-1}, \hat{X}_{-2}, \ldots, X_{-1}, X_{-2}, \ldots)\right],$$
$$H^k = E\left[-\log P(\hat{X}_0|\hat{X}_{-k}^{-1}, X_{-k}^{-1})\right],$$
$$\vec{P}^k(\hat{X}^N|X^N) = \vec{P}(\hat{X}^k|X^k) \prod_{i=k+1}^{N} P(\hat{X}_i|\hat{X}_{i-k}^{i-1}, X_{i-k}^{i-1}),$$
$$\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0) = \prod_{i=1}^{N} P(\hat{X}_i|\hat{X}_{-\infty}^{i-1}, X_{-\infty}^{i-1}).$$

Lemma A.2.

$$-\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N) \to H^k,$$
$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) \to \vec{H}^\infty(\hat{X}||X).$$

Proof.

$$-\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N) = -\frac{1}{N}\vec{P}(\hat{X}^k|X^k) - \frac{1}{N}\sum_{i=k+1}^{N} \log P(\hat{X}_i|\hat{X}^{i-1}_{i-k}, X^{i-1}_{i-k}) \tag{A.2}$$
$$\to 0 + H^k \quad \text{by the ergodic theorem.}$$

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) = -\frac{1}{N}\sum_{i=1}^{N} \log P(\hat{X}_i|, \hat{X}^{i-1}_{-\infty}, X^{i-1}_{-\infty}) \tag{A.3}$$
$$\to \vec{H}^\infty(\hat{X}||X) \quad \text{by the ergodic theorem.}$$

□

Lemma A.3.

$$H^k \to \vec{H}^\infty(\hat{X}||X), \quad \vec{H}(\hat{X}||X) = \vec{H}^\infty(\hat{X}||X).$$

Proof. We know that $H^k \to \vec{H}(\hat{X}||X)$, since the joint process is stationary and $\{H_k\}$ is a non-increasing sequence of non-negative numbers. So we only need to show that $H^k \to \vec{H}^\infty(\hat{X}||X)$. The Martingale convergence theorem says that

$$P(\hat{x}_0|\hat{X}^{-1}_{-k}, X^{-1}_{-k}) \to P(\hat{x}_0|\hat{X}^{-1}_{-\infty}, X^{-1}_{-\infty}).$$

Since $\hat{\mathcal{X}}$ is a finite alphabet and $p\log p$ is bounded, by the dominated convergence theorem,

$$\begin{aligned}
\lim_{k\to\infty} H^k &= \lim_{k\to\infty} E\left[-\sum_{\hat{x}_0\in\hat{\mathcal{X}}} P(\hat{x}_0|\hat{X}^{-1}_{-k}, X^{-1}_{-k}) \log P(\hat{x}_0|\hat{X}^{-1}_{-k}, X^{-1}_{-k})\right] \\
&= E\left[-\sum_{\hat{x}_0\in\hat{\mathcal{X}}} P(\hat{x}_0|\hat{X}^{-1}_{-\infty}, X^{-1}_{-\infty}) \log P(\hat{x}_0|\hat{X}^{-1}_{-\infty}, X^{-1}_{-\infty})\right] \\
&= \vec{H}^\infty(\hat{X}||X).
\end{aligned}$$

Thus $H^k \to \vec{H}^{\infty}(\hat{X}||X)$. □

Lemma A.4.

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} &\leq 0, \\ \limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})} &\leq 0, \end{aligned}$$

where

$$\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) \triangleq \prod_{i=1}^{N} P(\hat{X}_i|X^{i-1}_{-\infty}, \hat{X}^{i-1}_{-\infty}). \tag{A.4}$$

Proof.

(A.5)

$$\begin{aligned} E\left[\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)}\right] &= \sum_{\hat{x}^N, x^N} P(\hat{x}^N, x^N) \frac{\prod_{i=1}^{k} P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) \cdot \prod_{i=k+1}^{N} P(\hat{x}_i|\hat{x}^{i-1}_{i-k}, x^{i-1}_{i-k})}{\prod_{i=1}^{N} P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1})} \\ &= \sum_{\hat{x}^N, x^N} P(\hat{x}^k, x^k) \cdot \prod_{i=k+1}^{N} P(x_i|x^{i-1}, \hat{x}^i) P(\hat{x}_i|\hat{x}^{i-1}_{i-k}, x^{i-1}_{i-k}) = 1, \end{aligned}$$

where the last equality follows by evaluating the sum first over x_N, then over $\hat{x}_N$, then over x_{N-1} and so on. Using the above in Markov's inequality, we have

$$\Pr\left\{\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq N^2\right\} \leq \frac{1}{N^2} \tag{A.6}$$

or

$$\Pr\left\{\frac{1}{N}\log\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq \frac{1}{N}\log N^2\right\} \leq \frac{1}{N^2}. \tag{A.7}$$

Since $\sum_{N=1}^{\infty} \frac{1}{N^2} < \infty$, the Borel-Cantelli lemma says that, with probability 1, the event

$$\left\{\frac{1}{N}\log\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq \frac{1}{N}\log N^2\right\}$$

occurs only for finitely many N. Thus

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \leq 0 \quad \text{with probability 1.}$$

The second part of the lemma is proved in a similar manner. Using conditional expectations, we can write

(A.8)
$$E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})}\right] = E_{\hat{X}^0_{-\infty},X^0_{-\infty}}\left[E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})}\middle|\hat{X}^0_{-\infty},X^0_{-\infty}\right]\right].$$

The inner expectation can be written as

(A.9)
$$\begin{aligned} &E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})}\middle|\hat{X}^0_{-\infty},X^0_{-\infty}\right] \\ &= \sum_{\hat{x}^N,x^N} P(\hat{x}^N,x^N|\hat{X}^0_{-\infty},X^0_{-\infty})\frac{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1},x^{i-1})}{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1},x^{i-1},\hat{X}^0_{-\infty},X^0_{-\infty})} \\ &= \sum_{\hat{x}^N,x^N}\prod_{i=1}^N P(x_i|\hat{x}^i,x^{i-1},\hat{X}^0_{-\infty},X^0_{-\infty})P(\hat{x}_i|\hat{x}^{i-1},x^{i-1}) = 1, \end{aligned}$$

where the last equality is obtained by evaluating the sum first over x_N, then over $\hat{x}_N$, then over x_{N-1} and so on. Using the Borel-Cantelli lemma as in the previous part, we obtain

$$\limsup_{N\to\infty}\frac{1}{N}\log\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})} \le 0.$$

□

Proof of Lemma 2.4- AEP. We will show that the sequence of random variables $-\frac{1}{N}\log\vec{P}(\hat{X}^N|X^N)$ is sandwiched between the upper bound H^k and the lower bound $\vec{H}^\infty(\hat{X}||X)$ for all $k \ge 0$. From Lemma A.4, we have

(A.10)
$$\limsup_{N\to\infty}\frac{1}{N}\log\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \le 0.$$

Since the limit $\frac{1}{N}\log\vec{P}^k(\hat{X}^N|X^N)$ exists (Lemma A.2), we can write (A.10) as

(A.11)
$$\limsup_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}(\hat{X}^N|X^N)} \le \lim_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}^k(\hat{X}^N|X^N)} = H^k.$$

The second part of Lemma A.4 can be written as

(A.12)
$$\liminf_{N\to\infty}\frac{1}{N}\log\frac{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})}{\vec{P}(\hat{X}^N|X^N)} \ge 0.$$

Since the limit $\frac{1}{N}\log\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})$ exists (Lemma A.2), we can rewrite (A.12) as

$$\liminf_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}(\hat{X}^N|X^N)}\geq\lim_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}(\hat{X}^N|X^N_{-\infty},\hat{X}^0_{-\infty})}=\vec{H}^\infty(\hat{X}||X) \tag{A.13}$$

Combining (A.11) and (A.13), we have

(A.14)

$$\vec{H}^\infty(\hat{X}||X)\leq\liminf_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}(\hat{X}^N|X^N)}\leq\limsup_{N\to\infty}\frac{1}{N}\log\frac{1}{\vec{P}(\hat{X}^N|X^N)}\leq H^k\quad\text{for all } k.$$

By Lemma A.3, $H^k\to\vec{H}(\hat{X}||X)=\vec{H}^\infty(\hat{X}||X)$. Thus

$$\lim_{N\to\infty}-\frac{1}{N}\log\vec{P}(\hat{X}^N|X^N)=\vec{H}(\hat{X}||X). \tag{A.15}$$

□

A.2 Proof of Direct Part of Theorem 2

The approach we will take is as follows. We build a source code for the $X-F$ block in Figure 2.4 (Section 2.4), a system without feed-forward. Here, the code-functions themselves are considered 'reconstructions' of the source sequences. We will then connect the $X-F$ and the $X-\hat{X}$ systems to prove the achievability of $R_{ff}(D)$.

For the sake of clarity, we present the proof in two parts. The first part establishes the background for making the connection between the $X-F$ and $X-\hat{X}$ systems. In the second part, we will construct random codes for the system without feed-forward and show the achievability of $R_{ff}(D)$ using the results of the first part. We describe

the second part in detail for the probability of error criterion. The proof for the expected distortion case is omitted since it is similar.

Part I

Let $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^*_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be the sequence of distributions that achieves the infimum in Theorem 2. In this part, we wish to construct a joint distribution over $X^N, \hat{X}^N$ and F^N, say $Q^*_{F^N,X^N,\hat{X}^N}$, such that the marginal over X^N and $\hat{X}^N$ satisfies

$$Q^*_{X^N,\hat{X}^N} = P_{X^N} P^*_{\hat{X}^N|X^N}. \tag{A.16}$$

To do this, as will be shown in the sequel, the only distribution we can choose is the code-function distribution P_{F^N}. We pick P_{F^N} such that the induced distribution $Q^*_{F^N,X^N,\hat{X}^N}$ has certain desired properties and (A.16) is also satisfied.

For any N, the joint distribution $P_{X^N} P^*_{\hat{X}^N|X^N}$ can be split, as in (2.10), as

$$P_{X^N} P^*_{\hat{X}^N|X^N} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|\hat{X}^n,X^{n-1}}, \tag{A.17}$$

where the marginals, given by P^{ch} and P^{dec}, can be considered the fictitious test-channel from $\hat{X}$ to X and the set of 'decoder' distributions to this test-channel, respectively.

Let P_{F^N} be any distribution on the space of code-functions. Given P_{F^N} and the test channel P_{ch} in (A.17), we now define a joint distribution over $Q_{X^N,F^N,\hat{X}^N}$ over $(\mathcal{X}^N, \mathcal{F}^N, \hat{\mathcal{X}}^N)$, imposing the following constraints.

1. For $n = 1, \ldots, N$,

$$Q_{\hat{X}_n|F_n,X^{n-1}}(\hat{x}_n|f_n, x^{n-1}) = \begin{cases} 1, \text{ if } \hat{x}_n = f_n(x^{n-1}) \\ 0, \text{ otherwise.} \end{cases} \tag{A.18}$$

2.

$$Q_{F_n|F^{n-1},X^{n-1}}(f_n|f^{n-1}, x^{n-1}) = P_{F_n|F^{n-1}}(f_n|f^{n-1}) \qquad n = 1, \ldots, N. \tag{A.19}$$

3. For $\hat{x}^n = f^n(x^{n-1})$,

(A.20)
$$Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n,\hat{x}^n,x^{n-1}) = P^{ch}_{X_n|\hat{X}^n,X^{n-1}}(x_n|\hat{x}^n,x^{n-1}), \quad n=1,\ldots,N.$$

A joint distribution Q is said to be *nice* with respect to P_{F^N} and $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^N$ if $\forall x^N \in \mathcal{X}^N, f^N \in \mathcal{F}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$, the three constraints above hold. It is important to note that in general, for a given problem of source coding with feed-forward, the joint distribution on $X^N, F^N, \hat{X}^N$ induced from an arbitrary encoder-decoder pair does not satisfy these conditions. We just want to construct a joint distribution Q over the variables of interest satisfying the above conditions for the direct coding theorem.

Given a code-function distribution P_{F^N} and the test channel $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^N$, there exists a unique joint distribution $Q_{F^N,X^N,\hat{X}^N}$ that is nice with respect to them. This follows from the following arguments.

(A.21)
$$\begin{aligned} Q_{F^N,X^N,\hat{X}^N} &= \left\{\prod_{n=1}^N Q_{X_n|F^n,X^{n-1}} \cdot Q_{F_n|F^{n-1},X^{n-1}}\right\} \cdot Q_{\hat{X}^N|F^N,X^N} \\ &= \left\{\prod_{n=1}^N Q_{X_n|F^n,X^{n-1}} \cdot P_{F_n|F^{n-1}}\right\} \cdot \delta_{\hat{X}^N=F^N(X^{N-1})} \quad , \end{aligned}$$

where we have used (A.18) and (A.19) to obtain the second equality. Now we can use the fact that $\hat{x}_n = f_n(x^{n-1})$ to write

(A.22)
$$\begin{aligned} Q_{X_n|F^n,X^{n-1}}(x_n|f^n,x^{n-1}) &= Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n,\hat{x}^n,x^{n-1}) \\ &= P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\left(x_n|x^{n-1},f^n(x^{n-1})\right), \end{aligned}$$

where we have used (A.20) for the second equality. Thus the unique nice joint distribution is given by

(A.23)
$$\begin{aligned} &Q_{F^N,X^N,\hat{X}^N}(f^N,x^N,\hat{x}^N) \\ &= \prod_{n=1}^N P_{F_n|F^{n-1}}(f_n|f^{n-1}) \cdot \prod_{n=1}^N P^{ch}_{X_n|X^{n-1},\hat{X}^{n-1}}(x_n|f^n(x^{n-1}),x^{n-1}) \cdot \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}. \end{aligned}$$

Keeping P^{ch} fixed, (A.23) says that choosing P_{F^N} automatically determines a unique nice distribution. We want to choose P_{F^N} such that the resulting nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$ satisfies

$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}, \tag{A.24}$$

so that (A.16) is satisfied.

Definition A.5. For a test-channel $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^N$, we call a code-function distribution P_{F^N} 'good' with respect to a decoder distribution $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^N$ if the following holds for the nice induced distribution $Q_{F^N,X^N,\hat{X}^N}$:

$$\prod_{n=1}^{N} Q_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}) = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}), \tag{A.25}$$
$$\forall x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \hat{\mathcal{X}}^N.$$

This definition of 'good' is equivalent to, but slightly different from that in [77]. The next Lemma says that it is possible to find such a good P_{F^N}. For the sake of clarity, we give the proof although it is found in [77] in a different flavor.

Lemma A.6. *For a test-channel $\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\}_{n=1}^N$, there exists a code-function distribution P_{F^N} good with respect to a decoder distribution $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^N$.*

Proof. For all f^N, define for $n = 1, \ldots, N$

$$\text{graph}(f_n) \triangleq \{(x^{n-1}, \hat{x}_n) : f_n(x^{n-1}) = \hat{x}_n\} \subset \mathcal{X}^{n-1} \times \hat{\mathcal{X}} \tag{A.26}$$

(A.27)

$$P_{F_n|F^{n-1}}(f_n|f^{n-1}) \triangleq \prod_{(b^{n-1},a_n)\in \text{graph}(f_n)} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(a_n|f_1,\ldots,f_{n-1}(b^{n-2}),b^{n-1}).$$

We will show that $P_{F^N} = \prod_{n=1}^N P_{F_n|F^{n-1}}$ is good with respect to $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^\infty$. We give the proof in two parts. In part A, we obtain an expression for the induced

decoder distribution given P_{F^N} and P^{ch}. Part B of the proof uses this expression to show that (A.27) defines a good code-function distribution. Actually, we first need to show that for all n, $P_{F_n|F^{n-1}}$ defined above is a valid probability distribution. This part is omitted since it can be shown using arguments similar to those in Part B.

Part A

Define $\forall n \in \{1, \cdots, N\}$

$$\Gamma_n(x^{n-1}, \hat{x}_n) \triangleq \left\{ f_n : f_n(x^{n-1}) = \hat{x}_n \right\}, \tag{A.28}$$

$$\Gamma^n(x^{n-1}, \hat{x}^n) \triangleq \left\{ f^n : f_i(x^{i-1}) = \hat{x}_i, \quad i = 1, \ldots, n \right\}. \tag{A.29}$$

Given the test-channel $\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\}_{n=1}^N$ and a code function distribution P_{F^N}, a unique nice distribution $Q_{F^N,X^N,\hat{X}^N}$ is determined. We now show that the induced decoder distribution is given by

(A.30)

$$Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) = P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})\right), \ \forall n.$$

This is Lemma 5.2 in [77], but we repeat the proof here for the sake of completeness.

Note that $(\hat{x}^{n-1}, x^{n-1})$ uniquely determines $(\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1})$ and vice versa. Therefore,

$$Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) = Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}). \tag{A.31}$$

Now $(x^{n-1}, \hat{x}_n)$ uniquely determines $(\Gamma_n(x^{n-1}, \hat{x}_n), x^{n-1})$ and vice versa. Thus we must have

$$\begin{aligned} &Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}) \\ &\quad = Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}). \end{aligned} \tag{A.32}$$

Since Q is nice, it satisfies (A.19). Hence

$$\begin{aligned} &Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1}, x_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}) \\ &\quad = P_{F_N|F^{N-1}}(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})). \end{aligned} \tag{A.33}$$

Combining (A.31), (A.32) and (A.33), we obtain the expression in (A.30).

Part B

We now show that P_{F^N} defined by (A.27) is good with respect to P^{dec}. For a pair $x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \mathcal{X}^N$, consider

(A.34)

$$\begin{aligned}\sum_{f^N : f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N) &= P_{F^N}(\Gamma^N(x^{N-1}, \hat{x}^N)) \\ &= P_{F^N}\left(\Gamma_1(\hat{x}_1), \ldots, \Gamma_n(x^{n-1}, \hat{x}_n), \ldots, \Gamma_N(x^{N-1}, \hat{x}_N)\right) \\ &= \prod_{n=1}^{N} P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1}, \hat{x}_n) | \Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})\right).\end{aligned}$$

Substituting (A.30) in the above equation, we get

$$\sum_{f^N : f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N) = \prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n | x^{n-1}, \hat{x}^{n-1}). \tag{A.35}$$

We can also write

(A.36)

$$\begin{aligned}\sum_{\substack{f^N : \\ f^N(x^{N-1})=\hat{x}^N}} P_{F^N}(f^N) &= \sum_{f_1 : f_1=\hat{x}_1} \cdots \sum_{\substack{f_n : \\ f_n(x^{n-1})=\hat{x}_n}} \cdots \sum_{\substack{f_N : \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{n=1}^{N} P_{F_n|F^{n-1}}(f_n | f^{n-1}) \\ = \sum_{f_1 : f_1=\hat{x}_1} P_{F_1}(f_1) \cdots &\sum_{\substack{f_n : \\ f_n(x^{n-1})=\hat{x}_n}} P_{F_n|F^{n-1}}(f_n | f^{n-1}) \cdots \sum_{\substack{f_N : \\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N | f^{N-1}).\end{aligned}$$

We evaluate the Nth inner summation in the above equation as

(A.37)

$$\begin{aligned}
&\sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N|f^{N-1}) \\
&\qquad \overset{(a)}{=} \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{(b^{N-1},a_N)\in \mathrm{gr}(f_N)} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\ldots,f_{N-1}(b^{N-2}),b^{N-1}) \\
&\qquad = P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}) \\
&\qquad\quad \cdot \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{\substack{(b^{N-1},a_N)\in \mathrm{gr}(f_N) \\ b^{N-1}\neq x^{N-1}}} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\ldots,f_{N-1}(b^{N-2}),b^{N-1}) \\
&\qquad \overset{(b)}{=} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}) \\
&\qquad\quad \cdot \prod_{b^{N-1}\neq x^{N-1}} \sum_{a_N} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\ldots,f_{N-1}(b^{N-2}),b^{N-1}) \\
&\qquad = P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}),
\end{aligned}$$

where $f_1,\ldots,f_{N-1}$ are specified by the $N-1$ outer summations and 'gr' has been used as shorthand for graph. (a) follows from (A.27) and (b) follows from an observation similar to

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} xyz = \sum_{x\in\mathcal{X}} x \cdot \sum_{y\in\mathcal{Y}} y \cdot \sum_{z\in\mathcal{Z}} z.$$

Now, the $(N-1)$th inner sum in (A.36) can be shown to be equal to $P^{dec}_{\hat{X}_{N-1}|\hat{X}^{N-2},X^{N-2}}(\hat{x}_{N-1}|\hat{x}^{N-2},x^{N-2})$ in a similar fashion. Thus we can compute the summations in (A.36) sequentially from $n=N$ down to $n=1$. Substituting in (A.36), we get

$$\sum_{\substack{f^N: \\ f^N(x^{N-1})=\hat{x}^N}} P_{F^N}(f^N) = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}). \tag{A.38}$$

From (A.35) and (A.38), we have

(A.39)
$$\prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}) = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}) \qquad n=1,\ldots,N.$$

completing the proof of the lemma. □

To summarize, we have the following:

- The code-function distribution $P^*_{F^N}$ and the test-channel $\left\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\right\}_{n=1}^{N}$ determine a unique nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$ given by (A.23).
- For a test-channel $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^{N}$, we can find a code function distribution $P^*_{F^N}$ to be good with respect to P^{dec}, i.e., the set of induced 'decoder' distributions of Q^* satisfying the relation

(A.40)
$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}, \qquad n=1,\ldots,N.$$

Hence we have

(A.41)
$$\begin{aligned} Q^*_{X^N,\hat{X}^N} &= \prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot Q^*_{X_n|X^{n-1},\hat{X}^n} \\ &= \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|X^{n-1},\hat{X}^n} \\ &= P_{X^N} \cdot P^*_{\hat{X}^N|X^N}. \end{aligned}$$

Equation (A.41) is the key to connect the $X-F$ source code without feed-forward to the $X-\hat{X}$ code with feed-forward. We are now ready to prove Theorem 2.

Part II (The probability of error constraint)

For any N, pick $M = 2^{NR}$ N-length code-functions independently according to $P^*_{F^N}$. Denote this $(N, 2^{NR})$ codebook by $\mathcal{C}_N$. Define the "distortion" $d'_N(x^N, f^N) =$

$d\left(x^N, f^N(x^{N-1})\right)$. Let

$$A(\mathcal{C}_N) = \left\{x^N \in \mathcal{X}^N : \exists f^N \in \mathcal{C}_N \quad with \quad d'_N(x^N, f^N) \leq D + \delta\right\} \tag{A.42}$$

The set $A^c(\mathcal{C}_N)$ represents the set of x^N's that are not well represented by the chosen codebook. We will show that $P_{X^N}(A^c(\mathcal{C}_N))$, averaged over all realizations of $\mathcal{C}_N$, goes to 0 as $N \to \infty$ as long as $R > R_{ff}(D)$. Indeed,

$$\begin{aligned} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &= \sum_{\mathcal{C}_N} P^*_{F^N}(\mathcal{C}_N) \sum_{x^N \notin A(\mathcal{C}_N)} P_{X^N}(x^N) \\ &= \sum_{x^N} P_{X^N}(x^N) \sum_{\mathcal{C}_N : x^N \notin A(\mathcal{C}_N)} P^*_{F^N}(\mathcal{C}_N). \end{aligned} \tag{A.43}$$

The last sum on the right-hand side of (A.43) is the probability of choosing a codebook that does not represent the particular x^N with a distortion $D + \delta$. Define the set

$$B_{N,\delta} = \left\{(x^N, f^N) : d'_N(x^N, f^N) < \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta, \quad \frac{1}{N} i_{Q^*}(x^N; f^N) < \overline{I}_{\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X} \to X) + \delta\right\}, \tag{A.44}$$

where $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is as in Theorem 2, and $\overline{I}(\hat{X} \to X)$ is computed with the distribution $\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ and is therefore equal to $R_{ff}(D)$. Define an indicator function

$$K(x^N, f^N) = \begin{cases} 1, \text{ if } (x^N, f^N) \in B_{N,\delta} \\ 0, \text{ otherwise.} \end{cases} \tag{A.45}$$

We will also need the following Lemma, whose proof is given in Appendix A.2.1.

Lemma A.7 (a). $Q^*_{F^N|X^N}(f^N|x^N) \leq P^*_{F^N}(f^N) 2^{N[R_{ff}(D)+\delta]}$, $\forall (x^N, f^N) \in B_{N,\delta}$.

(b) $Q^*_{X^N,F^N}(B_{N,\delta}) \to 1$ *as* $N \to \infty$.

Since $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ achieves $R_{ff}(D)$ we have $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D$. Hence, for any f^N that does not represent a given x^N with distortion $\leq D + \delta$, the pair (x^N, f^N) does not belong to $B_{N,\delta}$. The probability that a code function chosen randomly according to $P^*_{F^N}$ does not represent a given x^N with distortion within $D + \delta$ is

$$\begin{aligned} P^*_{F^N}\left(d'_N(x^N, F^N) \geq D + \delta\right) &\leq P^*_{F^N}\left(K(x^N, F^N) = 0\right) \\ &= 1 - \sum_{f^N} P^*_{F^N}(f^n) K(x^N, f^N). \end{aligned} \tag{A.46}$$

Thus, the probability that none of 2^{NR} code functions, each independently chosen according to $P^*_{F^N}$, represent a given x^N with distortion $D+\delta$ is upper bounded by

$$\left(1-\sum_{f^N} P^*_{F^N}(f^N)K(x^N,f^N)\right)^{2^{NR}}.$$

This implies

$$E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] \le \sum_{x^N} P_{X^N}(x^N)\left(1-\sum_{f^N} P^*_{F^N}(f^n)K(x^N,f^N)\right)^{2^{NR}}$$

$$\le \sum_{x^N} P_{X^N}(x^N)\left(1-\exp_2\left\{-N(R_{ff}(D)+\delta)\right\}\sum_{f^N} Q^*_{F^N|X^N}(f^N|x^N)K(x^N,f^N)\right)^{2^{NR}}$$

where the last inequality follows from Lemma A.7(a). Using the inequality

$$(1-xy)^N \le 1-x+2^{-yN} \qquad \text{for } 0\le x,y\le 1,$$

in (A.47), we get

$$\begin{aligned} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &\le 1+\exp_2\left[-\exp_2[N(R-R_{ff}(D)-\delta)]\right] \\ &\quad -\sum_{x^N,f^N} P_{X^N}(x^N)Q^*_{F^N|X^N}(f^N|x^N)K(x^N,f^N) \\ &= 1-Q^*_{F^N,X^N}(B_{N,\delta})+\exp_2\Big(-\exp_2[N(R-R_{ff}(D)-\delta)]\Big). \end{aligned} \tag{A.47}$$

When $R > R_{ff}(D)+\delta$, using part(b) of Lemma A.7, we have

$$\lim_{N\to\infty} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right]=0. \tag{A.48}$$

Therefore, there exists at least one sequence of codes $\{\mathcal{C}_N\}$ such that

$$\limsup_{N\to\infty} P_{X^N}\left(A^c(\mathcal{C}_N)\right)=0.$$

In other words, there exists a sequence of codebooks $\{\mathcal{C}_N\}$ of code-functions for which

$$\lim_{N\to\infty}\Pr\left\{x^N\in\mathcal{X}^N:\quad d_N(x^N,f^N(x^{N-1}))>D+\delta,\qquad \forall f^N\in\mathcal{C}_N\right\}=0. \tag{A.49}$$

The theorem follows. □

A.2.1 Proof of Lemma A.7

Proof: **(a)** From the definition, we have

$$i_{Q^*_{X^N,F^N}}(x^N;f^N) = \log \frac{Q^*_{F^N|X^N}(f^N|x^N)}{Q^*_{F^N}(f^N)}$$

Therefore,

$$\begin{aligned} Q^*_{F^N|X^N}(f^N|x^N) &= Q^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N;f^N)] \\ &= P^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N;f^N)], \end{aligned} \tag{A.50}$$

where the second equality follows because $P^*_{F^N}$ is used to construct Q^*. Moreover,

$$\frac{1}{N} i_{Q^*}(x^N;f^N) < \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X) + \delta, \qquad \forall (x^n,f^N)\in B_{N,\delta}. \tag{A.51}$$

Substituting the above in (A.50), we get part (a) of the lemma.

(b) The code function distribution $P^*_{F^N}$, the test-channel $\left\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\right\}_{n=1}^N$ determines a nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$, given by (A.23). Under these conditions Q^* satisfies

$$\begin{aligned} \frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} &= \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^N}}{Q^*_{X^N}} \\ &\overset{a}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^n}}{Q^*_{X^N}} \\ &\overset{b}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},\hat{X}^n}}{Q^*_{X^N}} \\ &= \frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}}, \end{aligned} \tag{A.52}$$

where, as before, $\vec{Q}^*_{\hat{X}^N|X^N} = \prod_{n=1}^N Q^*_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}$. (a) holds because the condition $Q_{X_n|X^{n-1},F^N} = Q_{X_n|X^{n-1},F^n}$ is equivalent to (A.19). This is shown in [54] as a condition for a channel not to have have feedback. (b) follows from (A.19) and (A.20). (A.52) is essentially Lemma 5.1 in [77]. Thus we have

(A.53)

$$i_{Q^*_{X^N,F^N}}(f^N;x^N) = \frac{1}{N}\log\frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} = \frac{1}{N}\log\frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}} = \vec{i}_{Q^*_{\hat{X}^N,X^N}}(\hat{x}^N;x^N).$$

Define

$$\vec{P}^{dec}_{\hat{X}^N|X^N} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}},$$
$$\vec{P}^{ch}_{X^N|\hat{X}^N} = \prod_{n=1}^{N} P^{ch}_{X_n|X^{n-1},\hat{X}^n}.$$

Since $P^*_{F^N}$ is chosen to be good with respect to $\vec{P}^{dec}_{\hat{X}^N|X^N}$ for the test channel P^{ch}, we have from (A.41)

$$Q^*_{X^N,\hat{X}^N} = \vec{Q^*}_{\hat{X}^N|X^N}\vec{Q^*}_{X^N|\hat{X}^N} = \vec{P}^{dec}_{\hat{X}^N|X^N}\vec{P}^{ch}_{X^N|\hat{X}^N} = P_{X^N}P^*_{\hat{X}^N|X^N}. \tag{A.54}$$

Using (A.54) in (A.53), we get

$$i_{Q^*_{X^N,F^N}}(f^N;x^N) = \vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N;x^N). \tag{A.55}$$

Now,

$$\begin{aligned} Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta}) &= Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad d'_N(x^N,f^N) \ge \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta \\ &\qquad\qquad \text{or} \quad \frac{1}{N} i_{Q^*}(x^N;f^N) \ge \overline{I}_{\mathbf{P}_{\mathbf{X}}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X) + \delta\Big) \\ &\le Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad d'_N(x^N,f^N) \ge \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) \\ &\quad + Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad \frac{1}{N} i_{Q^*}(x^N;f^N) \ge \overline{I}_{\mathbf{P}_{\mathbf{X}}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X) + \delta\Big). \end{aligned} \tag{A.56}$$

Since

$$d'_N(x^N,f^N) = d_N(x^N,f^N(x^{N-1})) = d_N(x^N,\hat{x}^N), \tag{A.57}$$

the first term in (A.56) equals

$$\begin{aligned} &Q^*_{X^N,\hat{X}^N}\Big((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \ge \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) \\ &\qquad = P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \ge \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big), \end{aligned} \tag{A.58}$$

where we have used (A.54). Since $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is the lim sup in probability of $d_N(x^N,\hat{x}^N)$,

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \ge \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) = 0. \tag{A.59}$$

The second term in ((A.56)) equals

(A.60)

$$Q^*_{F^N,X^N,\hat{X}^N}\left((f^N,x^N,\hat{x}^N):\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N)\geq \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X)+\delta\right)$$
$$=P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N):\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N)\geq \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X)+\delta\right),$$

where we have used (A.55) for the first equality and (A.54) for the next. Since $\overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X)$ is the $\limsup_{\text{in prob}}$ of $\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N;x^N)$,

(A.61)

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N):\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N)\geq \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X)+\delta\right)=0.$$

Equations (A.59) and (A.61) imply

(A.62)
$$\lim_{N\to\infty} Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta})=0,$$

proving part (b) of the lemma. □

A.3 Proof of Converse Part of Theorem 2

Let $\{\mathcal{C}_N\}_{N=1}^{\infty}$ be any sequence of codes, with rate R, that achieve distortion D (either expected distortion D or probability-1 distortion D depending on the criterion used). For any given block length N, there is an induced $P_{F^N|X^N}$. (equal to 1 for the code function f^N chosen to represent X^N and 0 for the other $2^{NR}-1$ code functions). This, along with the source distribution $P(X^N)$, determines P_{F^N}, a 2^{NR}-point discrete distribution. Thus, given the source distribution and the encoding and decoding rules, a joint distribution is induced. $\forall x^N\in\mathcal{X}^N, \hat{x}^N\in\hat{\mathcal{X}}^N, f^N\in\{f^N[i], i=1,\cdots,2^{NR}\}$, the induced distribution is given by

(A.63)
$$\hat{Q}_{X^N,F^N,\hat{X}^N}(x^N,f^N,\hat{x}^N)=P_{X^N}(x^N)\cdot P_{F^N|X^N}(f^N|x^N)\cdot\delta_{\{\hat{x}^N=f^N(x^{N-1})\}}.$$

All probability distributions in the remainder of this section are marginals drawn from the induced joint distribution in (A.63). We first show that for any such induced distribution, we have

$$\overline{H}(F) = \limsup_{inprob} \frac{1}{N} \log \frac{1}{P(F^N)} \leq R. \tag{A.64}$$

Equivalently, we show that for any $\delta > 0$,

$$\lim_{N\to\infty} \Pr\left(\frac{1}{N} \log \frac{1}{P(F^N)} > R + \delta\right) = 0. \tag{A.65}$$

We have

$$\begin{aligned}
\Pr\left(\frac{1}{N} \log \frac{1}{P(F^N)} > R + \delta\right) &= \Pr\left(P(F^N) < 2^{-N(R+\delta)}\right) \\
&= \sum_{f^N : 0 < P_{F^N}(f^N) < 2^{-N(R+\delta)}} P_{F^N}(f^N) \\
&\leq \sum_{f^N : P_{F^N}(f^N) > 0} 2^{-N(R+\delta)} \\
&= 2^{NR} \cdot 2^{-N(R+\delta)} \\
&= 2^{-N\delta} \to 0 \quad \text{as} \quad N \to \infty,
\end{aligned} \tag{A.66}$$

thereby proving (A.64). Thus we have

$$R \geq \overline{H}(F) \geq \overline{H}(F) - \underline{H}(F|X) \geq \overline{I}(F;X), \tag{A.67}$$

where the last inequality follows from Lemma 2 in [76]. We need the following lemma, whose proof is found in Section A.3.1.

Lemma A.8. *For any sequence of codes as defined above, we have*

$$\overline{I}(F;X) \geq \overline{I}(\hat{X} \to X), \tag{A.68}$$

where the above quantities are computed with joint distribution induced by the code.

Using this lemma in (A.67), we obtain

$$R \geq \overline{I}(\hat{X} \to X). \tag{A.69}$$

By assumption, the sequence of codes with rate R achieves distortion D. This means that the induced output distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ satisfies the distortion constraint in Theorem 2. Therefore, we have

$$R \geq \overline{I}(\hat{X} \to X) \geq R_{ff}(D). \tag{A.70}$$ □

A.3.1 Proof of Lemma A.8

Let $\hat{Q}_{X^N,F^N,\hat{X}^N}$ be the joint distribution induced by the source code as in (A.63). From Definition 2.12, we have

$$i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N) = \log \frac{P(X^N|F^N)}{\vec{P}(X^N|\hat{X}^N)} = \log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)}, \tag{A.71}$$

where the distributions are those induced from the source code. The upper-case notation we have used indicates that we want to consider the probabilities and the information quantities as random variables. We will first show that

$$\liminf_{inprob} \frac{1}{N}\left(i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N)\right) \geq 0. \tag{A.72}$$

This is equivalent to proving that for any $\delta > 0$,

$$\lim_{N\to\infty} P\left(\frac{1}{N}\log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) = 0. \tag{A.73}$$

Since $F^n(X^{n-1}) = \hat{X}^n$, we have

$$P(X^N|F^N) = \prod_{n=1}^N P(X_n|X^{n-1},F^N) = \prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n). \tag{A.74}$$

Therefore,

$$
\begin{aligned}
& P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) \\
& = P\left(\prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n) < 2^{-N\delta}\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)\right) \qquad \text{(A.75)} \\
& = \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N),
\end{aligned}
$$

where

$$
\mathcal{G} = \left\{(f^N,x^N,\hat{x}^N) : \prod_{n=1}^N P_{X_n|X^{n-1},F^N,\hat{X}^n}(x_n|x^{n-1},f^N,\hat{x}^n) < 2^{-N\delta}\prod_{n=1}^N P_{X_n|X^{n-1},\hat{X}^n}(x_n|x^{n-1},\hat{x}^n)\right\}.
$$

In the remainder of this section, we drop the subscripts of the probabilities since the arguments make it clear what P refers to in each case.

(A.76)

$$
\begin{aligned}
\sum_{\mathcal{G}}\hat{Q}_{F^N,X^N,\hat{X}^N}(f^N,x^N,\hat{x}^N) &= \sum_{\mathcal{G}} P(f^N)P(x^N,\hat{x}^N|f^N) \\
&= \sum_{\mathcal{G}} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n,f^N)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\
&< 2^{-N\delta}\sum_{\mathcal{G}} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\
&\le 2^{-N\delta}\sum_{x^N,f^N,\hat{x}^N} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\
&\overset{(a)}{=} 2^{-N\delta}\sum_{f^N} P(f^N)\sum_{(x^N,\hat{x}^N):f^N(x^{N-1})=\hat{x}^N}\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\
&\overset{(b)}{=} 2^{-N\delta}\sum_{f^N} P(f^N)\sum_{x^N}\prod_{n=1}^N P(x_n|x^{n-1},f^n(x^{n-1})) \\
&\overset{(c)}{=} 2^{-N\delta}\cdot 1,
\end{aligned}
$$

where (a) follows from the fact that $\hat{x}^N = f^N(x^{N-1})$ and (b) since the term $P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N)$ is equal to 1 when $\hat{x}_n = f_n(x^{n-1})$ and zero otherwise. (c)

is obtained by evaluating the inner summation first over x_N, then over x_{N-1} and observing that all the f_n's are constant in the inner summation. Therefore (A.75) becomes

$$\begin{aligned} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) &= \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N) \\ &< 2^{-N\delta}. \end{aligned} \tag{A.77}$$

Hence

$$\lim_{N\to\infty} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) = 0. \tag{A.78}$$

Thus we have proved (A.72). Now, using the inequality

$$\liminf_{inprob}(a_n+b_n) \le \limsup_{inprob} a_n + \liminf_{inprob} b_n \tag{A.79}$$

in (A.72), we get

$$\begin{aligned} 0 \le \liminf_{inprob}\frac{1}{N}\left(i(F^N;X^N)-\vec{i}(\hat{X}^N;X^N)\right) &\le \limsup_{inprob}\frac{1}{N}i(F^N;X^N) + \liminf_{inprob} -\frac{1}{N}\vec{i}(\hat{X}^N;X^N) \\ &= \limsup_{inprob}\frac{1}{N}i(F^N;X^N) - \limsup_{inprob}\frac{1}{N}\vec{i}(\hat{X}^N;X^N). \end{aligned} \tag{A.80}$$

Or,

$$\overline{I}(F;X) \ge \overline{I}(\hat{X}\to X), \tag{A.81}$$

completing the proof of Lemma A.8. □

A.4 Proof of Theorem 3

The source distribution is a sequence of distributions $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^\infty$, where for each n, P_{X^n} is a product distribution. The rate-distortion function for an arbitrary memoryless source without feed-forward is

$$R_{DMS}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\le D} \overline{I}(\hat{X};X), \tag{A.82}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^{N} d_i(X_i, \hat{X}_i)]. \tag{A.83}$$

Part 1: We first show that for a memoryless distortion measure with an expected distortion constraint, a memoryless conditional distribution achieves the infimum. Let $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be any conditional distribution, for which the sup-directed information is $\overline{I}(\hat{X};X)$ and expected distortion is D. We will show that there exists a memoryless conditional distribution $\mathbf{P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}}$ such that $\overline{I}_{ML}(\hat{X};X) \leq \overline{I}(\hat{X};X)$ and the expected distortion with $\mathbf{P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}}$ is the same, i.e., D. From the corresponding joint distribution $\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{X^n,\hat{X}^n}\}$, form a memoryless joint distribution $\mathbf{P}_{\mathbf{X}}\mathbf{P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^{ML}_{X^n,\hat{X}^n}\}$ as follows. Set

$$P^{ML}_{X^n,\hat{X}^n} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i}, \tag{A.84}$$

where $P_{X_i,\hat{X}_i}, i \in \{1,\cdots,n\}$ are the marginals of $P_{X^n,\hat{X}^n}$. Clearly, for any N, the expected distortion with $P_{X^N,\hat{X}^N}$

$$E_{P_{X^N,\hat{X}^N}}[\frac{1}{N}\sum_{i=1}^{N} d_i(X_i,\hat{X}_i)] = \frac{1}{N}\sum_{i=1}^{N} E_{P_{X_i,\hat{X}_i}} d_i(X_i,\hat{X}_i) \tag{A.85}$$

is the same for $P^{ML}_{X^N,\hat{X}^N}$. We need to show

$$\overline{I}_{ML}(\hat{X};X) \leq \overline{I}(\hat{X};X) \qquad \text{or}$$

$$\limsup_{inprob} \frac{1}{N} i_{ML}(\hat{X}^N;X^N) \leq \limsup_{inprob} \frac{1}{N} i(\hat{X}^N;X^N).$$

To prove that

$$\limsup_{inprob} a_n \geq \limsup_{inprob} b_n, \tag{A.86}$$

it is enough to show that $\liminf_{inprob} a_n - b_n \geq 0$. This would imply

$$\begin{aligned} 0 \leq \liminf_{inprob} a_n - b_n &\leq \limsup_{inprob} a_n + \liminf_{inprob} -b_n \\ &= \limsup_{inprob} a_n - \limsup_{inprob} b_n. \end{aligned} \tag{A.87}$$

We have

(A.88)

$$\frac{1}{N}\left(i(\hat{X}^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)=\frac{1}{N}\log\frac{P(\hat{X}^N,X^N)}{P(\hat{X}^N)\prod_{i=1}^N P(X_i)}\cdot\prod_{i=1}^N\frac{P(X_i)P(\hat{X}_i)}{P(X_i,\hat{X}_i)}$$
$$=\frac{1}{N}\log\frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^N P(X_i|\hat{X}_i)}.$$

We want to show that the $\liminf_{inprob}$ of the expression in (A.88) is ≥ 0. This is equivalent to showing that for any $\delta>0$,

$$\lim_{N\to\infty}\Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)<-\delta\right]=0. \tag{A.89}$$

Let

$$\mathcal{G}=\left\{(x^N,\hat{x}^N):P_{X^N|\hat{X}^N}(x^N|\hat{x}^N)<2^{-N\delta}\prod_{i=1}^N P_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\right\}.$$

Then,

(A.90)

$$\Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)<-\delta\right]=\Pr\left[\frac{1}{N}\log\frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^N P(X_i|\hat{X}_i)}<-\delta\right]$$
$$=\Pr\left[P(X^N|\hat{X}^N)<2^{-N\delta}\prod_{i=1}^N P(X_i|\hat{X}_i)\right]=\sum_{(x^N,\hat{x}^N)\in\mathcal{G}}P_{\hat{X}^N}(\hat{x}^N)P_{X^N|\hat{X}^N}(x^N|\hat{x}^N)$$
$$\overset{(a)}{\leq}2^{-N\delta}\sum_{(x^N,\hat{x}^N)\in\mathcal{G}}\prod_{i=1}^N P_{\hat{X}_i|X^{i-1}}(\hat{x}_i|x^{i-1})P_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\overset{(b)}{=}2^{-N\delta}\cdot 1,$$

where (a) follows from the definition of $\mathcal{G}$ and (b) is obtained by evaluating the sum first over x_N, then over $\hat{x}_N$ and so on. The arguments in (A.86) and (A.87) complete the proof that the infimum achieving distribution can be assumed to be memoryless in source coding without feed-forward. We now show that feed-forward does not change the rate-distortion function of the memoryless source.

Part 2: Let $\{\mathcal{C}_N\}_{N=1}^{\infty}$ be any sequence of codes with feed-forward, with rate R, that is achievable at distortion D. For any given block length N, a joint distribution

described by (A.63) is induced:

$$\hat{Q}_{X^N,F^N,\hat{X}^N} = P_{X^N} \cdot P_{F^N|X^N} \cdot \delta_{\{\hat{X}^N = F^N(X^{N-1})\}}. \tag{A.91}$$

All probability distributions in the remainder of this section are marginals drawn from this induced joint distribution. As in Part 1, define a memoryless distribution $\hat{Q}^{ML}_{X^N,\hat{X}^N} \triangleq \prod_{n=1}^{N} \hat{Q}_{X_n,\hat{X}_n}$. The subscript ML on an information quantity will imply that $\hat{Q}^{ML}_{X^N,\hat{X}^N}$ is the distribution being used to compute it. As shown in Appendix A.3 ((A.64) to (A.67)), for this joint distribution we have

$$R \geq \overline{H}(F) \geq \overline{H}(F) - \underline{H}(F|X) \geq \overline{I}(F;X). \tag{A.92}$$

It remains to show that when the source is memoryless,

$$\begin{aligned} \overline{I}(F;X) &\geq \overline{I}_{ML}(\hat{X};X) \quad \text{or} \\ \limsup_{inprob} \frac{1}{N} i(F^N;X^N) &\geq \limsup_{inprob} \frac{1}{N} i_{ML}(\hat{X}^N;X^N). \end{aligned} \tag{A.93}$$

As in Part 1 of this proof, it suffices to show that

$$\liminf_{inprob} \frac{1}{N} \left(i(F^N;X^N) - i(\hat{X}^N;X^N) \right) \geq 0$$

or equivalently that for all $\delta > 0$,

$$\lim_{N\to\infty} \Pr\left[\frac{1}{N} \left(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N) \right) < -\delta \right] = 0. \tag{A.94}$$

Noting that $\hat{Q}^{ML}_{X^N,\hat{X}^N}$ is memoryless, we have

(A.95)

$$\begin{aligned} \frac{1}{N} \left(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N) \right) &= \frac{1}{N} \log \frac{\hat{Q}(F^N,X^N)}{\hat{Q}(F^N) \prod_{n=1}^{N} P(X_n)} \cdot \prod_{n=1}^{N} \frac{P(X_n)\hat{Q}(\hat{X}_i)}{\hat{Q}(X_n,\hat{X}_n)} \\ &= \frac{1}{N} \log \frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^{N} \hat{Q}(X_n|\hat{X}_n)}. \end{aligned}$$

Hence, we have

(A.96)

$$\begin{aligned}
&\Pr\left[\frac{1}{N}\left(i(F^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)<-\delta\right]=\Pr\left[\frac{1}{N}\log\frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^N\hat{Q}(X_n|\hat{X}_n)}<-\delta\right]\\
&\quad=\Pr\left[\hat{Q}(X^N|F^N)<2^{-N\delta}\prod_{n=1}^N\hat{Q}(X_n|\hat{X}_n)\right]\\
&\quad=\Pr\left[(x^N,f^N,\hat{x}^N):\hat{Q}_{X^N|F^N}(x^N|f^N)<2^{-N\delta}\prod_{n=1}^N\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\\
&\quad=\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\hat{Q}_{X^N|F^N}(x^N|f^N)\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N)\\
&\quad\le 2^{-N\delta}\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^N\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N),
\end{aligned}$$

where

$$\nu(f^N)\triangleq\left\{(x^N,\hat{x}^N):\hat{Q}_{X^N|F^N}(x^N|f^N)<2^{-N\delta}\prod_{i=1}^N\hat{Q}_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\right\}.$$

Since f^N and x^N determine the reconstruction $\hat{x}^N$, $\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N)=1$ if $\hat{x}^N=f^N(x^{N-1})$ and 0 otherwise. Thus we have

(A.97)

$$\begin{aligned}
&\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^N\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N)\\
&\qquad=\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{x^N}\prod_{n=1}^N\hat{Q}_{X_n|\hat{X}_n}(x_n|f_n(x^{n-1}))=1,
\end{aligned}$$

where the inner summation is computed first over x_N, then x_{N-1} and so on up to x_1. Thus

(A.98)

$$\begin{aligned}
\Pr\left[\frac{1}{N}\left(i(F^N;X^N)-i(\hat{X}^N;X^N)<-\delta\right)\right]&\le 2^{-N\delta}\\
&\to 0\quad\text{as}\quad N\to\infty,
\end{aligned}$$

proving (A.94). We have shown that any achievable rate R (with feed-forward) satisfies

$$R \geq \overline{I}_{ML}(\hat{X}; X).$$

This implies that the rate-distortion function with feed-forward is the same as that without feed-forward. □

A.5 Proof of Theorems 4 and 5

We first give the proof of Theorem 4. We will use the error-exponent result proved by Iriyama for a general source without feed-forward. For source coding without feed-forward, Theorem 1 in [40] gives the formula for the minimum achievable rate with error exponent r:

$$\sup_{\mathbf{Y}: D_l(\mathbf{Y}||\mathbf{X})<r} R^*(D|\mathbf{Y}) \leq R(D, r|\mathbf{X}) \leq \sup_{\mathbf{Y}: D_l(\mathbf{Y}||\mathbf{X})\leq r} R^*(D|\mathbf{Y}), \tag{A.99}$$

with equalities if $R(D, r|\mathbf{X})$ is continuous at r. In (A.99), the quantities have the same definitions as those in Section 2.5, except that there is no feed-forward.

Recall from Section 2.4.1 that every source coding system with feed-forward is equivalent to a source coding system without feed-forward defined in terms of code-functions. For the no-feed-forward version, the reconstruction is a code-function F^N and the distortion is given by

$$d_n(X^n, F^n) = d_n(X^n, F^n(X^{n-1})), \quad \forall n.$$

Hence, (A.99) holds for the source coding problem with source X and reconstruction F. Clearly, any rate-distortion function for the no-feed-forward $X - F$ system is the

same as the rate-distortion function for the system $X - \hat{X}$ with feed-forward. Thus we obtain (2.58).

To prove the second part of Theorem 4, we use Theorem 5 from [40]. Applying this theorem to the $X - F$ source coding system (no feed-forward), we obtain

$$\text{(A.100)} \qquad \inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D} \underline{I}(F;\mathbf{Y}) \leq R_{ff}^{*}(D|\mathbf{Y}) \leq \inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D_1} \underline{I}(F;\mathbf{Y}), \quad 0 < D_1 < D$$

We can use the same procedure used in Appendix A.2 to prove the direct part of Theorem 2 to show that

$$\inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D} \underline{I}(\mathbf{F};\mathbf{Y}) = \inf_{\hat{\mathbf{Y}}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\leq D} \underline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}),$$

completing the proof.

Theorem 5 can be proved in a similar fashion, using code-functions and appealing to Theorems 2 and 4 in [40].

A.6 Proof of (2.69)

$$
\begin{aligned}
I_k(\hat{X}^N \to X^N) &= I(\hat{X}^N; X^N) - \sum_{n=k+1}^{N} I(X^{n-k}; \hat{X}_n | \hat{X}^{n-1}) \\
&= E\left[\log \frac{P(x^N, \hat{x}^N)}{P(x^N)P(\hat{x}^N)}\right] - \sum_{n=k+1}^{N} E\left[\log \frac{P(x^{n-k}, \hat{x}_n|\hat{x}^{n-1})}{P(x^{n-k}|\hat{x}^{n-1})P(\hat{x}_n|\hat{x}^{n-1})}\right] \\
&= E\left[\log \frac{P(x^N, \hat{x}^N)}{P(x^N)P(\hat{x}^N)}\right] - E\left[\log \frac{\prod_{n=k+1}^{N} P(\hat{x}_n|x^{n-k}, \hat{x}^{n-1})}{\prod_{n=k+1}^{N} P(\hat{x}_n|\hat{x}^{n-1})}\right] \\
&= E\left[\log \frac{P(x^N, \hat{x}^N)}{P(x^N)P(\hat{x}^N)}\right] - E\left[\log \frac{\prod_{n=k+1}^{N} P(\hat{x}_n|x^{n-k}, \hat{x}^{n-1}) \cdot P(\hat{x}^k)}{P(\hat{x}^N)}\right] \\
&= E\left[\log \frac{P(x^N, \hat{x}^N)}{P(x^N) \cdot P(\hat{x}^k) \prod_{n=k+1}^{N} P(\hat{x}_n|x^{n-k}, \hat{x}^{n-1})}\right] \\
&= E\left[\log \frac{P(x^N, \hat{x}_{k+1}^N|\hat{x}^k)}{P(x^N) \prod_{n=k+1}^{N} P(\hat{x}_n|x^{n-k}, \hat{x}^{n-1})}\right] \\
&= E\left[\log \frac{\prod_{n=k+1}^{N} P(\hat{x}_n, x_{n-k}|\hat{x}^{n-1}, x^{n-k-1}) \cdot P(x_{N-k+1}^N|\hat{x}^N, x^{N-k})}{P(x^N) \prod_{n=k+1}^{N} P(\hat{x}_n|x^{n-k}, \hat{x}^{n-1})}\right] \\
&= E\left[\log \frac{\prod_{n=k+1}^{N} P(x_{n-k}|\hat{x}^n, x^{n-k-1}) \cdot P(x_{N-k+1}^N|\hat{x}^N, x^{N-k})}{P(x^N)}\right] \\
&= E\left[\log \frac{\prod_{m=1}^{N-k} P(x_m|\hat{x}^{m+k-1}, x^{m-1}) \cdot \prod_{m=N-k+1}^{N} P(x_m|\hat{x}^N, x^{m-1})}{\prod_{m=1}^{N-k} P(x_m|x^{m-1}) \cdot \prod_{m=N-k+1}^{N} P(x_m|x^{m-1})}\right] \\
&= E\left[\log \prod_{m=1}^{N} \frac{P(x_m|\hat{x}^{m+k-1}, x^{m-1})}{P(x_m|x^{m-1})}\right] \\
&= \sum_{m=1}^{N} I(\hat{X}^{m+k-1}; X_m|X^{m-1}).
\end{aligned}
$$

APPENDIX B

Proofs for Chapter 3

B.1 Proof of Lemma 3.3

$$
\begin{aligned}
&I_k(X^N \to Y^N) \\
&= I(X^N;Y^N) - \sum_{n=k+1}^{N} I(Y^{n-k};X_n|X^{n-1}) \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)P(y^N)} \\
&\qquad - \sum_{n=k+1}^{N}\sum_{x^n,y^{n-k}} P(x^n,y^{n-k})\log\frac{P(x_n,y^{n-k}|x^{n-1})}{P(x_n|x^{n-1})P(y^{n-k}|x^{n-1})} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)\cdot P(y^N)} - \sum_{n=k+1}^{N}\sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x_n|x^{n-1},y^{n-k})}{P(x_n|x^{n-1})} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)\cdot P(y^N)} - \sum_{x^N,y^N} P(x^N,y^N)\log\prod_{n=k+1}^{N}\frac{P(x_n|x^{n-1},y^{n-k})}{P(x_n|x^{n-1})} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{\prod_{n=k+1}^{N}P(x_n|x^{n-1},y^{n-k})\cdot P(y^N)} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{\vec{P}^k(x^N|y^N)\cdot P(y^N)}.
\end{aligned}
$$

B.2 Proof of Theorem 9

B.2.1 The joint distribution

The input distribution is of the form

$$\mathbf{P}^{k,l}_{\mathbf{X}|\mathbf{Y},\mathbf{S}} = \{P_{X_n|X^{n-1},Y^{n-k},S^{n-l}}\}_{n=1}^{\infty}. \tag{B.1}$$

In order to specify the joint distribution of all the variables in the system, we first define the notion of a code-function in the context of our problem.

Definition B.1. A channel code-function f^N is a sequence of N mappings $\{f_n\}_{n=1}^N$ such that $f_n : \mathcal{Y}^{n-k} \times \mathcal{S}^{n-l} \to \mathcal{X}$.

Figure B.1 depicts the role of a code-function. At time 0, we can imagine an outer encoder mapping the message to a code-function F^N. Thus a channel codebook of block length N, rate R is actually a set of 2^{NR} code-functions. In each future time instant, this code-function F^N determines the actual channel input using the feedback and state-information available at that time. The dashed box in Figure B.1 is a 'channel' without feedback with input F^N and output (Y^N, S^N).

The different variables in the systems (with n denoting time) are F_n, S_n, X_n, Y_n. For our analysis, we will introduce another variable $E_n \triangleq (Y^{n-k}, S^{n-l})$. In other words, E_n represents the information available at the encoder at time n. We can now describe the joint distribution of the system variables, $Q(F^N, S^N, X^N, Y^N, E^N)$. The channel is given by (3.15) and the channel input distribution

$$\{P_{X_n|X^{n-1},Y^{n-k},S^{n-l}}\}_{n=1}^{\infty} = \{P_{X_n|X^{n-1},E^{n-1}}\}_{n=1}^{\infty}$$

is given by (B.1). Recall that $F^N = \{F_n\}_{n=1}^N$ is chosen at time 0. Also note that the state information S^N is generated independently of the channel input and output. Further, without loss of generality, we can assume that S^N is generated at time 0.

This assumption does not alter the joint distribution even if the symbols S_n are actually generated in real time[82]. Thus we have

(B.2)

$$Q(F^N, S^N, X^N, Y^N, E^N) = P(F^N)P(S^N)\prod_{n=1}^{N} Q(X_n, Y_n, E_n|X^{n-1}, Y^{n-1}, E^{n-1}, F^N, S^N)$$

$$= P(F^N)P(S^N)\prod_{n=1}^{N} Q(X_n|X^{n-1}, Y^{n-1}, E^{n-1}, F^N, S^N)Q(Y_n|X^n, Y^{n-1}, E^{n-1}, F^N, S^N)$$

$$\cdot Q(E_n|X^n, Y^n, E^{n-1}, F^N, S^N)$$

$$\stackrel{(a)}{=} P(F^N)\cdot P(S^N)\cdot\prod_{n=1}^{N} P(X_n|X^{n-1}, E^{n-1})P^{ch}(Y_n|X^n, Y^{n-1}, S^n)\delta_{(E_n=(Y^{n-k},S^{n-l}))},$$

$$\stackrel{(b)}{=} P(F^N)\cdot P(S^N)\prod_{n=1}^{N}\delta_{X_n=F_n(X^{n-1},E^{n-1})}P^{ch}(Y_n|X^n, Y^{n-1}, S^n)\delta_{(E_n=(Y^{n-k},S^{n-l}))}$$

where $(a), (b)$ are justified as follows. $(X^{n-1}, Y^{n-1}, E^{n-1}, F^N, S^N)$ represents all the events that have occured until time $n-1$. From this information, the encoder is only allowed to choose the input X_n depending on (X^{n-1}, E^{n-1}), the information it has access to. Hence

$$Q(X_n|X^{n-1}, Y^{n-1}, E^{n-1}, F^N, S^N) = P(X_n|X^{n-1}, E^{n-1}). \tag{B.3}$$

From the definition of the channel, given all the information until time n, the output Y_n depends only on (X^n, Y^{n-1}, S^n). So we have

$$Q(Y_n|X^n, Y^{n-1}, E^{n-1}, F^N, S^N) = P^{ch}(Y_n|X^n, Y^{n-1}, S^n). \tag{B.4}$$

Finally, (b) follows from the definition of code-function.

Since we have not made any assumptions such as stationarity and ergodicity of the distributions involved, we will need to use information spectrum quantities to

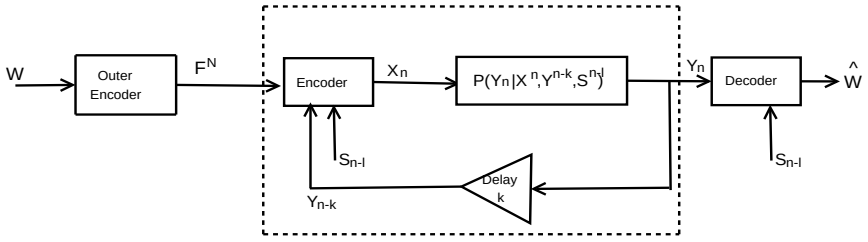

Figure B.1: The role of a code-function

characterize the capacity. The following definitions will be required.

$$\vec{P}^1_{X^N||Y^N|S^N}(x^N||y^N|s^N) \triangleq \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^N}(x_n|x^{n-1},y^{n-1},s^N), \tag{B.5}$$

$$\vec{i}(x^N;y^N|s^N) \triangleq \log \frac{P_{X^N,Y^N,S^N}(x^N,y^N,s^N)}{\vec{P}^1_{X^N||Y^N|S^N}(x^N||y^N|s^N)P_{Y^N}(y^N)} \quad , \tag{B.6}$$

$$\underline{I}(X \to Y|S) \triangleq \liminf_{inprob} \frac{1}{N}\vec{i}(X^N;Y^N|S^N). \tag{B.7}$$

B.2.2 Proof of capacity result- Direct part

We first characterize the input distribution $Q_{X_n|X^{n-1},E^{n-1}}$ induced by a given code-function distribution P_{F^N}. Define the following:

$$\begin{aligned} \text{graph}(f_n) &= \left\{(e^{n-1},x_n) : f_n(e^{n-1}) = x_n\right\}, \\ \Gamma_n(e^{n-1},x^n) &= \left\{f^n : f_j(e^{j-1}) = x_j, j = 1,\cdots,n\right\}. \end{aligned} \tag{B.8}$$

Lemma B.2. *Suppose we are given a channel* $\{P_{Y_n|X^n,Y^{n-1},S^n}\}_{n=1}^N$*, state distribution* $\{P_{S^n}\}_{n=1}^N$ *and a code-function distribution* P_{F^N}*, such that the joint distribution is given by* Q *as in* (B.2)*. Then the induced input distribution for all* n *is given by*

$$Q(x_n|x^{n-1},y^{n-1},s^N) = Q(x_n||x^{n-1},e^{n-1}) = P_{f_n|f^{n-1}}(\Gamma_n(e^{n-1},a_n)|\Gamma^{n-1}(e^{n-2},a^{n-2})).$$

Proof. The first equality follows from the arguments used to justify (B.2). The second inequality can be proved using steps similar to Lemma 5.1 in [78].

Lemma B.3. *For any valid joint distribution on the system as in* (B.2)*,we have*

$$\frac{Q_{F^N,Y^N}(f^N,y^N)}{Q_{F^N}(f^N)Q_{Y^N}(y^N)} = \frac{Q_{X^N,Y^N|S^N}(x^N,y^N|s^N)}{\vec{Q}^1_{X^N||Y^N|S^N}(x^N||y^N|s^N)Q_{Y^N}(y^N|s^N)}.$$

Proof. Similar to Lemma 5.2 in [78].

As a consequence of the above lemma, we have for our system

$$\underline{I}(F;Y) = \underline{I}(X \to Y|S). \tag{B.9}$$

Outline: The 'channel' from F^N to Y^N is one without feedback. Hence its capacity is characterized by $\underline{I}(F;Y)$ [84]. Let $\{P^*_{X_n|X^{n-1},E^{n-1}}\}_{n=1}^{\infty}$ be the input distribution that maximizes $\underline{I}(X \to Y|S)$. Suppose we choose a code function distribution such that for every N, its induced input distribution (given by Lemma B.2) is equal to $P^*_{X_n|X^{n-1},E^{n-1}}$. Then, by virtue of (B.9), we have $\underline{I}(F;Y) = \underline{I}_{P^*}(X \to Y|S)$. Since rates arbitrarily close to $\underline{I}(F;Y)$ are achievable, we obtain the direct part. Thus the key to proving the direct part is choosing a suitable a code function distribution $\{P_{F^N}\}$. This is done as follows.

Definition B.4. Given an input distribution $\{P_{X_n|X^{n-1},E^{n-1}}\}_{n=1}^{N}$, a code-function distribution P_{F^N} is called 'good' with respect to the input distribution if it induces a joint distribution Q (according to (B.2)) that satisfies the following for $\forall x^n, y^n, s^n, \quad n = 1, \ldots, N$:

$$Q(x_n|x^{n-1}, e^{n-1}) = P(x_n|x^{n-1}, e^{n-1}).$$

Lemma B.5. *Given any input distribution $\{P_{X_n|X^{n-1},E^{n-1}}\}_{n=1}^{N}$, there exists a code-function distribution P_{F^N} that is good with respect to it.*

Proof. Define $P_{F^N} = \prod_{n=1}^{N} P_{F_n|F^{n-1}}$ as

$$P(f_n|f^{n-1}) = \prod_{(e^{n-1},a_n)\in \mathrm{graph}(f_n)} P_{X_n|X^{n-1},E^{n-1}}(a_n|f_1, f_2(e_1), \ldots, f_{n-1}(e^{n-2}), e^{n-1}), \quad \forall n.$$

It can be verified that $P_{F_n|F^{n-1}}$ is a valid probability distribution and that

$$P_{F_n|F^{n-1}}(\Gamma_n(e^{n-1}, a_n)|f^{n-1}) = P_{X_n|X^{n-1},E^{n-1}}(x_n|f^{n-1}(e^{n-2}), e^{n-1}).$$

Then the goodness of P_{F^N} is established using Lemma B.2.

The proof of the direct part can be completed as described in the outline above.

B.2.3 Converse part

As in [78], the converse is a generalization of the converse found in [84]. We first have the following lemma.

Lemma B.6. *Every* $(N, 2^{NR})$ *channel code with probability of error* ϵ *satisfies* $\forall \gamma > 0$

$$\epsilon \geq Q\left((x^N, y^N, s^N): \quad \frac{1}{N}\frac{Q_{X^N,Y^N|S^N}(x^N, y^N|s^N)}{\vec{Q}^1_{X^N||Y^N|S^N}(x^N||y^N|s^N)Q_{Y^N}(y^N|s^N)} \leq R - \gamma\right) - 2^{-\gamma N}.$$

Proof. First use Theorem 4 of [84] for the '$F-Y$' channel without feedback and then Lemma B.3. The converse can then be proved in a similar fashion to Theorem 5.2 in [84].

APPENDIX C

Proofs for Chapter 4

C.1 Proof of Theorem 11

To keep the notation in the proof simple, we will use P to denote the source distribution $\mathbf{P_X}$ and W to denote the conditional distribution $\mathbf{P_{\hat{X}|X}}$. Let $\tilde{W}$ denote any other conditional distribution $\tilde{\mathbf{P}}_{\mathbf{\hat{X}|X}}$ such that

$$\limsup_{in\ prob\ P\tilde{W}} d_n(X^n, \hat{X}^n) \leq \limsup_{in\ prob\ PW} d_n(X^n, \hat{X}^n).$$

In the various stages of the proof, the distribution used to calculate the information quantity is indicated in the subscript. A ˜ indicates that the quantity is calculated using the joint distribution $P\tilde{W}$.We will show that if (4.7) is satisfied, then

$$\overline{I}_{k\ P\tilde{W}}(\hat{X} \to X) \geq \overline{I}_{k\ PW}(\hat{X} \to X), \tag{C.1}$$

thus proving the optimality of W.

Step 1: We will first show that

$$\begin{aligned} \overline{I}_{k\ P\tilde{W}}(\hat{X} \to X) &\triangleq \limsup_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{\tilde{W}_{\hat{X}^n|X^n}}{\vec{\tilde{P}}^k_{\hat{X}^n|X^n}} \\ &\geq \limsup_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}. \end{aligned} \tag{C.2}$$

Since

$$\limsup_{inprob} a_n - \limsup_{inprob} b_n \leq \limsup_{inprob}(a_n - b_n),$$

to show (C.2), it is enough to show that

$$\limsup_{in\ prob\, P\tilde{W}} \frac{1}{n} \log \frac{P_{X^n} W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n} \cdot P_{X^n}} - \frac{1}{n} \log \frac{P_{X^N} \tilde{W}_{\hat{X}^n|X^n}}{\vec{\tilde{P}}^k_{\hat{X}^n|X^n} \cdot P_{X^n}} \leq 0. \tag{C.3}$$

Noting that $P_{X^N} W_{\hat{X}^n|X^n} = P_{X^n,\hat{X}^n}$, the LHS of (C.3) is

$$\begin{aligned}
\limsup_{in\ prob\, P\tilde{W}} \frac{1}{n} \log \frac{\frac{P_{X^n,\hat{X}^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \cdot \vec{\tilde{P}}^k_{\hat{X}^n|X^n}}{P_{X^n} \tilde{W}_{\hat{X}^n|X^n}} &= - \liminf_{in\ prob\, P\tilde{W}} \frac{1}{n} \log \frac{P_{X^n} \tilde{W}_{\hat{X}^n|X^n}}{\frac{P_{X^n,\hat{X}^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \cdot \vec{\tilde{P}}^k_{\hat{X}^n|X^n}} \\
&= - \liminf_{in\ prob\, P\tilde{W}} \frac{1}{n} \log \frac{P_{X^n} \tilde{W}_{\hat{X}^n|X^n}}{G_{\hat{X}^n,X^n}} \\
&\overset{(a)}{\leq} 0,
\end{aligned}$$

where $G_{\hat{X}^n,X^n}$ is the joint distribution on $(X^N, \hat{X}^N)$ given by $\frac{P_{\hat{X}^n,X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \cdot \vec{\tilde{P}}^k_{\hat{X}^n|X^n}$ and (a) is shown to be true in Theorem 8(a) in [84]. Thus (C.2) holds.

Step 2: We now show that if

$$\limsup_{in\ prob\, P\tilde{W}} d_n(X^n, \hat{X}^n) \leq \limsup_{in\ prob\, PW} d_n(X^n, \hat{X}^n) \tag{C.4}$$

then

$$\limsup_{in\ prob\, P\tilde{W}} d_n(X^n, \hat{X}^n) + \beta_n(X^n) \leq \limsup_{in\ prob\, PW} d_n(X^n, \hat{X}^n) + \beta_n(X^n) \tag{C.5}$$

for an arbitrary sequence of functions $\beta_n(.)$. Intuitively, this seems reasonable since the distribution on X^n is the same for both LHS and RHS of (C.4) and so the inequality is preserved by adding any function of X^n. This can be formally proved as follows.

Recall that $\limsup_{in\ prob\, PW} a_n$ is the smallest number α such that

$$\lim_{n\to\infty} P_{PW}(a_n > \alpha) = 0.$$

Let

$$\limsup_{in\,prob\,PW} d(X^n, \hat{X}^n) = a_1 \quad \text{and} \quad \limsup_{in\,prob\,P\tilde{W}} d(X^n, \hat{X}^n) = a_2,$$

with $a_2 \leq a_1$. Also, let

$$\limsup_{in\,prob\,PW} d(X^n, \hat{X}^n) + \beta_n(X^n) = b_1.$$

Now

$$\begin{aligned}
&\lim_{n\to\infty} P_{PW}(d(X^n, \hat{X}^n) + \beta_n(X^n) > b_1) \\
&\qquad = \lim_{n\to\infty} \sum_{x^n, \hat{x}^n : d(x^n,\hat{x}^n) > b_1 - \beta_n(x^n)} P_{X^n}(x^n) W_{\hat{X}^n|X^n}(\hat{x}^n|x^n) \\
&\qquad = 0.
\end{aligned}$$

Since a_1 is the smallest number a such that

$$\lim_{n\to\infty} \sum_{x^n, \hat{x}^n : d(x^n,\hat{x}^n) > a} P_{X^n}(x^n) W_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = 0,$$

we have

$$\liminf_{n\to\infty} b_1 - \beta(x^n) \geq a_1 \text{ w.p.1 under } \mathbf{P_X}. \tag{C.6}$$

To prove (C.5), it suffices to show that for any $\alpha \geq b_1$,

$$\limsup_{n\to\infty} P_{P\tilde{W}}(d(X^n, \hat{X}^n) + \beta_n(X^n) > \alpha) = 0 \tag{C.7}$$

Consider (C.7) for any $\alpha \geq b_1$.

(C.8)

$$\begin{aligned}
&\limsup_{n\to\infty} P_{P\tilde{W}}(d(X^n, \hat{X}^n) + \beta_n(X^n) > \alpha) \leq \limsup_{n\to\infty} P_{P\tilde{W}}(d(X^n, \hat{X}^n) + \beta_n(X^n) > b_1) \\
&\qquad = \limsup_{n\to\infty} \sum_{x^n, \hat{x}^n : d(x^n,\hat{x}^n) > b_1 - \beta(x^n)} P_{X^n}(x^n) \tilde{W}_{\hat{X}^n|X^n}(\hat{x}^n|x^n) \\
&\qquad \overset{(a)}{\leq} \limsup_{n\to\infty} \sum_{x^n, \hat{x}^n : d(x^n,\hat{x}^n) > a_1} P_{X^n}(x^n) \tilde{W}_{\hat{X}^n|X^n}(\hat{x}^n|x^n) \\
&\qquad \overset{(b)}{\leq} \limsup_{n\to\infty} \sum_{x^n, \hat{x}^n : d(x^n,\hat{x}^n) > a_2} P_{X^n}(x^n) \tilde{W}_{\hat{X}^n|X^n}(\hat{x}^n|x^n) \\
&\qquad = 0,
\end{aligned}$$

where (a) holds due to (C.6) and (b) holds because $a_2 \leq a_1$.Thus (C.7) holds if (C.4) is true.

Step 3: Here we use the results of Steps 1 and 2 to prove (C.1). For $\lambda > 0$, we have

(C.9)

$$\begin{aligned}\bar{I}_{k\,P\tilde{W}}(\hat{X} \to X) - \bar{I}_{k\,PW}(\hat{X} \to X) &\overset{(a)}{\geq} \limsup_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} - \limsup_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \\ &\overset{(b)}{\geq} \limsup_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} - \limsup_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \\ &\quad + \limsup_{in\ prob\ P\tilde{W}} \left(\lambda d_n(X^n, \hat{X}^n) + \beta_n(X^n)\right) - \limsup_{in\ prob\ PW} \left(\lambda d_n(X^n, \hat{X}^n) + \beta_n(X^n)\right),\end{aligned}$$

where (a) follows from (C.2) in Step 1, (b) is from (C.4) in Step 2. Now we set

$$\lambda d_n(X^n, \hat{X}^n) + \beta_n(X^n) = -\frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}, \quad \lambda > 0 \tag{C.10}$$

to obtain

(C.11)

$$\begin{aligned}\bar{I}_{k\,P\tilde{W}}(\hat{X} \to X) - \bar{I}_{k\,PW}(\hat{X} \to X) &\overset{(a)}{\geq} \limsup_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} - \liminf_{in\ prob\ P\tilde{W}} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} \\ &\quad + \liminf_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} - \limsup_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}\end{aligned}$$

Now, since $\mathbf{P_X W_{\hat{X}|X}}$ is k-directed information stable, we have

$$\bar{I}_{k\,PW}(\hat{X} \to X) \triangleq \limsup_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} = \limsup_{N\to\infty} I_k(\hat{X}^N \to X^N). \tag{C.12}$$

Further, as a consequence of (4.6),

(C.13)

$$\bar{I}_{k\,PW}(\hat{X} \to X) = \limsup_{N\to\infty} I_k(\hat{X}^N \to X^N) = \liminf_{N\to\infty} I_k(\hat{X}^N \to X^N) = \underline{I}(\hat{X} \to X)$$

and so

$$\limsup_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}} = \liminf_{in\ prob\ PW} \frac{1}{n} \log \frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}$$

Using this in (C.11), we get

(C.14)

$$\bar{I}_{k\,P\tilde{W}}(\hat{X}\to X)-\bar{I}_{k\,PW}(\hat{X}\to X)\geq \limsup_{in\ prob\ P\tilde{W}}\frac{1}{n}\log\frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}-\liminf_{in\ prob\ P\tilde{W}}\frac{1}{n}\log\frac{W_{\hat{X}^n|X^n}}{\vec{P}^k_{\hat{X}^n|X^n}}$$

$$\geq 0.$$

Rearranging (C.10), we obtain the required condition.

C.2 Proof of Corollary 4.2

From Theorem 11, we know that for sufficiently large n, the distortion measure has to be of the form

$$\begin{aligned} d_n(x^n,\hat{x}^n) &= -c\cdot\frac{1}{n}\log\frac{P(x^n,\hat{x}^n)}{\vec{P}^k(\hat{x}^n|x^n)}+d_0(x^n)\\ &= -c\cdot\frac{1}{n}\sum_{i=1}^{n}\log\frac{P(x_i,\hat{x}_i|x^{i-1},\hat{x}^{i-1})}{P(\hat{x}_i|\hat{x}^{i-1},x^{i-k})}+d_0(x^n), \end{aligned} \tag{C.15}$$

where the last equality is obtained by writing $P(x^n,\hat{x}^n)=\prod_{i=1}^n P(x_i,\hat{x}_i|x^{i-1},\hat{x}^{i-1})$ and using (4.1) for $\vec{P}^k(\hat{x}^n|x^n)$. Since the joint distribution satisfies (4.9), we have

$$P(x_i,\hat{x}_i|x^{i-1},\hat{x}^{i-1})=P(x_i,\hat{x}_i|x_{i-m}^{i-1}). \tag{C.16}$$

$P(\hat{x}_i|\hat{x}^{i-1},x^{i-k})$ can be simplified as follows.

$$\begin{aligned} P(\hat{x}_i|\hat{x}^{i-1},x^{i-k}) &= \frac{P(\hat{x}^i,x^{i-k})}{P(\hat{x}^{i-1},x^{i-k})}\\ &= \frac{\sum_{x_{i-k+1}^i}P(\hat{x}^i,x^i)}{\sum_{\hat{x}_i,x_{i-k+1}^i}P(\hat{x}^i,x^i)}\\ &= \frac{\sum_{x_{i-k+1}^i}P(\hat{x}_{i-k+1}^i,x_{i-k+1}^i|\hat{x}^{i-k},x^{i-k})\cdot P(\hat{x}^{i-k},x^{i-k})}{\sum_{\hat{x}_i,x_{i-k+1}^i}P(\hat{x}_{i-k+1}^i,x_{i-k+1}^i|\hat{x}^{i-k},x^{i-k})\cdot P(\hat{x}^{i-k},x^{i-k})}\\ &\overset{(a)}{=} \frac{\sum_{x_{i-k+1}^i}P(\hat{x}_{i-k+1}^i,x_{i-k+1}^i|x_{i-k+1-m}^{i-k})}{\sum_{\hat{x}_i,x_{i-k+1}^i}P(\hat{x}_{i-k+1}^i,x_{i-k+1}^i|x_{i-k+1-m}^{i-k})}\\ &= \frac{P(\hat{x}_{i-k+1}^i|x_{i-k+1-m}^{i-k})}{P(\hat{x}_{i-k+1}^{i-1}|x_{i-k+1-m}^{i-k})}\\ &= P(\hat{x}_i|\hat{x}^{i-1},x_{i-k+1-m}^{i-k}). \end{aligned} \tag{C.17}$$

In the above chain of equalities, (a) has been obtained using the Markov structure of the joint distribution, given by (4.9). Substituting (C.16) and (C.17) in (C.15), we obtain the required expression.

C.3 Proof of Theorem 12

Let $\mathbf{P}_{\mathbf{X}|\mathbf{Y}}^{\prime k}$ denote any other input distribution that achieves lower cost than $\mathbf{P}_{\mathbf{X}|\mathbf{Y}}^{k}$ over the channel. In the sequel, the symbol $'$ will denote that the quantity is computed using the joint distribution

$$P'_{XY} \triangleq \vec{P}_{X|Y}^{\prime k} \cdot \vec{P}_{Y|X}^{ch}.$$

We will also use the notation

$$P_{XY} \triangleq \vec{P}_{X|Y}^{k} \cdot \vec{P}_{Y|X}^{ch}.$$

We have

$$\limsup_{in\ prob\, P'_{XY}} c_n(X^n, Y^n) \leq \limsup_{in\ prob\, P'_{XY}} c_n(X^n, Y^n). \tag{C.18}$$

We will show that if (C.18) is satisfied, then

$$\underline{I}_{P_{XY}}(X \to Y) \geq \underline{I}_{P'_{XY}}(X \to Y) \tag{C.19}$$

under the conditions of the theorem, thus proving the optimality of $\mathbf{P}_{\mathbf{X}|\mathbf{Y}}^{k}$.

Step 1: We will first show that

$$\begin{aligned} \underline{I}_{P'_{XY}}(X \to Y) &\triangleq \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}_{Y^n|X^n}^{ch}}{P'_{Y^n}} \\ &\leq \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}_{Y^n|X^n}^{ch}}{P_{Y^n}}. \end{aligned} \tag{C.20}$$

Since

$$\liminf_{inprob} a_n - \liminf_{inprob} b_n \geq \liminf_{inprob}(a_n - b_n),$$

to show (C.20), it is enough to see that

$$\liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{P'_{Y^n}}{P_{Y^n}} \geq 0$$

which is true from Theorem 8(a) in [84].

Step 2: Here we use Step 1 to prove (C.19). We have

(C.21)

$$\begin{aligned}
\underline{I}_{P_{XY}}(X \to Y) - \underline{I}_{P'_{XY}}(X \to Y) &\overset{(a)}{\geq} \liminf_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} - \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} \\
&\overset{(b)}{\geq} \liminf_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} - \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} \\
&\quad + \limsup_{in\ prob\, P'_{XY}} [\beta \cdot c_n(X^n, Y^n) + b_0] - \limsup_{in\ prob\, P_{XY}} [\beta \cdot c_n(X^n, Y^n) + b_0]
\end{aligned}$$

where $\beta > 0$ and (a) follows from (C.20) in Step 1, (b) is from (C.18). Now set

$$\beta c_n(X^n, Y^n) + b_0 = \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)}. \tag{C.22}$$

Since $\mathbf{P_{XY}}$ is directed information stable, we have from (4.16)

$$\underline{I}_{P_{XY}}(X \to Y) \triangleq \liminf_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} = \liminf_{N\to\infty} I(X^N \to Y^N). \tag{C.23}$$

Further, as a consequence of equality in (4.16),

(C.24)

$$\bar{I}_{P_{XY}}(X \to Y) = \limsup_{N\to\infty} I(X^N \to Y^N) = \liminf_{N\to\infty} I(X^N \to Y^N) = \underline{I}(X \to Y)$$

and so

$$\limsup_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} = \liminf_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}$$

Hence (C.21) becomes

(C.25)

$$\begin{aligned}
\underline{I}_{P_{XY}}(X \to Y) - \underline{I}_{P'_{XY}}(X \to Y) &\geq \limsup_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} - \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} \\
&\geq 0.
\end{aligned}$$

Rearranging (C.22), we get the result.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] R. Ahlswede. The rate-distortion region for multiple descriptions without excess rate. *IEEE Trans. Inf. Theory*, IT-31(6):721–726, Nov. 1985.

[2] P. H. Algoet and T. M. Cover. A Sandwich Proof of the Shannon-McMillan-Breiman Theorem. *The Annals of Probability*, 16(2):899–909, April 1988.

[3] V Ananthram. A large deviations approach to error exponents in source coding and hypothesis testing. *IEEE Transactions on Information Theory*, IT-4(4):938–943, July 1990.

[4] R. J. Barron, B. C., and G. W. Wornell. The Duality between Information Embedding and Source Coding with Side Information and Some Applications. *IEEE Transactions on Information Theory*, 49(5):1159–1180, May 2003.

[5] T. Berger. *Rate-distortion theory: A mathematical basis for data compression*. Prentice-Hall, Massachusetts, 1971.

[6] T. Berger and Z. Zhang. Minimum breakdown degradation in binary source encoding. *IEEE Trans. Inf. Theory*, IT-29(6):807–814, Nov. 1983.

[7] R. E. Blahut. Computation of channel capacity and rate-distortion function. *IEEE Trans. Inf. Theory*, 18(4):460–473, July 1972.

[8] R. E Blahut. *Principles and Practice of Information Theory*. Addison Wesley, Massachusetts, 1988.

[9] S.I. Bross and A. Lapidoth. An improved achievable region for the discrete memoryless two-user multiple-access channel with noiseless feedback. *IEEE Trans Inf Theory*, IT-51(3):811–833, March 2005.

[10] M. V. Burnashev. Data transmission over a discrete channel with feedback, random transmission time. *Problemy Perdachi Informatsii*, 12(4).

[11] P. E. Caines and C. W. Chan. Feedback between stationary processes. *IEEE Transactions on Automatic Control*, AC-20(378):498–508, 1975.

[12] J. Chen and T. Berger. The capacity of finite-state markov channels with feedback. *IEEE Transactions on Information Theory*, 51(3):780–798, 2005.

[13] T. Cover and M. Chiang. Duality Between channel capacity and rate-distortion with two-sided state information. *IEEE Trans. on Information Theory*, 48(6):1629–1638, June 2002.

[14] T. Cover, A. El Gamal, and M. Salehi. Multiple access channels with arbitrarily correlated sources. *IEEE Trans. Inf. Theory*, 26(6):648–657, November 1980.

[15] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[16] T. M. Cover and A. El Gamal. Achievable rates for multiple descriptions. *IEEE Trans. Inf. Theory*, IT-28(6):851–857, Nov 1982.

[17] T. M. Cover and S. K. Leung. An achievable rate region for multiple-access channel with feedback. *IEEE Trans Inf Theory*, IT-27:292–298, March 1981.

[18] T. M. Cover and S. Pombra. Gaussian feedback capacity. *IEEE Transactions on Information Theory*, IT-35:37–43, January 1989.

[19] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press,, New York, 1981.

[20] R. L. Dobrushin. General formulation of shannon's basic theorems of information theory. *AMS Translations, Series* 2, 33(3):323–438, 1961.

[21] R. L. Dobrushin. An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback. *Problemy Kibernetiki*, 8:161–168, 1962.

[22] S. C. Draper and A. Sahai. Variable-length channel coding with noisy feedback. *European Transactions on Telecommunications*, 19(4):355–370, June 2008.

[23] G. Dueck. Partial feedback for two-way and broadcast channels. *Information and Control*, IT-51(1):1–15, July 1980.

[24] F. Fu and R. W. Yeung. On the rate-distortion region for multiple descriptions. *IEEE Trans. Inf. Theory*, 48(7):2012–2021, July 2002.

[25] Jr. G. D. Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, IT-14:206–220, March 1968.

[26] N. T. Gaarder and J. K. Wolf. The capacity region of a multiple access discrete memoryless channel can increase with feedback. *IEEE Trans Inf Theory*, IT-21:100–102, 1975.

[27] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley and Sons, New York, 1968.

[28] A. El Gamal. The feedback capacity of degraded broadcast channels. *IEEE Trans. Inf. Theory*, IT-24(1):379–381, May 1978.

[29] M. Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: lossy source-channel communication revisited. *IEEE Transactions on Information Theory*, 49(5):1147–1158, 2003.

[30] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, June 1982.

[31] R. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[32] R. Gray, D. Neuhoff, and J. Omura. Process definitions of distortion-rate functions and source coding theorems. *IEEE Trans. Inform. Theory*, 21(5):524– 532, Sept. 1975.

[33] T. Han and S. Verdú. Approximation theory of output statistics. *IEEE Transactions on Information Theory*, 39(3):752–772, May 1993.

[34] T. S. Han. An information-spectrum approach to source coding theorems with a fidelity criterion. *IEEE Transactions on Information Theory*, 43(4):1145–1164, 1997.

[35] T. S. Han. The reliability functions of the general source with fixed-length coding. *IEEE Transactions on Information Theory*, 46(6):2117–2132, 2000.

[36] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer, 2002.

[37] T. S. Han. Joint source-channel coding revisited: Information-spectrum approach. *CoRR*, http://arxiv.org/abs/0712.2959, 2007.

[38] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Trans Inf Theory*, 12:448–455, October 1966.

[39] K. Iriyama. Probability of error for the fixed-length source coding of general sources. *IEEE Transactions on Information Theory*, 47(4):2466–2473, 2001.

[40] K. Iriyama. Probability of error for the fixed-length lossy coding of general sources. *IEEE Transactions on Information Theory*, 51(4):1498–1507, 2005.

[41] K. Iriyama and S. Ihara. The error exponent and minimum achievable rates for the fixed-length coding of general sources. *IEICE Trans. on Fund. of Electronics, Communications and Computer Sciences*, E84-A(10):1537–1543, 2001.

[42] J. M. Kahn, R. H. Katz, and K. S. J. Pister. Mobile networking for smart dust. *ACM/IEEE International Conference on Mobile Computing*, Seattle, WA, August 1999.

[43] R. L. Kashyap. Feedback coding schemes for an additive noise channel with a noisy feedback link. *IEEE Transactions on Information Theory*, 14(3):471–480, May 1968.

[44] Y. H. Kim. The feedback capacity of the first-order moving average gaussian channel. *IEEE Trans Info Theory*, IT-52(7):3063–3079, July 2006.

[45] Y. H. Kim. A coding theorem for a class of stationary channels with feedback. *IEEE Trans Info Theory*, IT-25(4):1488–1499, April 2008.

[46] G. Kramer. *Directed Information for channels with Feedback*. PhD thesis, Swiss Federal Institute of Technology, Zurich, 1998.

[47] G. Kramer. Capacity Results for the Discrete Memoryless Network. *IEEE Transactions on Information Theory*, 49(1):4–20, January 2003.

[48] S. I. Krich. Coding for a Delay-Dependent Fidelity Criterion. *IEEE Transactions on Information Theory*, pages 77–85, January 1974.

[49] G Longo. On the Error Exponent for Markov Sources. In *Proc. of the 2nd IEEE International symposium on Informtaion Theory (ISIT), Tsakhadsor, Soviet Union.* Budapest: Akademiai Kiado, 1971.

[50] H. Marko. The Bidirectional Communication Theory- A Generalization of Information Theory. *IEEE Trans. on Communications*, COM-21(12):1345–1351, December 1973.

[51] E. Martinian and G. W. Wornell. Source Coding with Fixed Lag Side Information. *Proceedings of the 42nd Annual Allerton Conference (Monticello, IL)*, 2004.

[52] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, IT-20:197–199, March 1974.

[53] J. Massey. Causality, Feedback and Directed Information. *Proceedings of the 1990 Symposium on Information Theory and its Applications (ISITA-90)*, pages 303–305, 1990.

[54] J. L. Massey. Network Information Theory- Some Tentative Definitions. DIMACS Workshop on Network Information Theory, April 2003.

[55] S Natarajan. Large deviations, hypotheses testing, and source coding for finite markov chains. *IEEE Transactions on Information Theory*, IT-31(3):360–365, May 1985.

[56] D. Neuhoff and R. Gilbert. Causal source codes. *IEEE Trans. Inform. Theory*, pages 701–713, Sept. 1982.

[57] Jim K Omura. A coding theorem for discrete time sources. *IEEE Transactions on Information Theory*, IT-19(4):490–498, July 1973.

[58] L. H. Ozarow. On the source coding problem with two channels and three receivers. *Bell Syst. Tech. J.*, 59(10):1909–1922, Dec 1980.

[59] L. H. Ozarow. The capacity of the white gaussian multiple access channel with feedback. *IEEE Transactions on Information Theory*, 30(4):623–628, 1984.

[60] L. H. Ozarow and S. K. Leung-Yan-Cheong. An achievable region and outer bound for the gaussian broadcast channel with feedback. *IEEE Transactions on Information Theory*, 30(4):667–671, 1984.

[61] K. R. Parthasarathy. On the integral representation of the rate of transmission of a stationary channel. *Ill. J. Math.*, 4:299–305, 1961.

[62] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. San Francisco, CA:Holden-Day, 1964. Translated by A.Feinstein.

[63] S. S. Pradhan. Approximation of test channels in source coding. In *Proc. of Conf. on Inform. Systems and Sciences (CISS)*, March 2004.

[64] S. S. Pradhan. Source coding with feedforward: Gaussian sources. In *Proc. of IEEE International Symposium on Information Theory*, page 212, June 2004.

[65] S. S. Pradhan. On the role of feedforward in gaussian sources: Point-to-point source coding and multiple description source coding. *IEEE Trans. Inf. Theory*, 53(1):331–349, January 2007.

[66] S. S. Pradhan, J. Chou, and K. Ramchandran. Duality between source coding and channel coding and its extension to the side-information case. *IEEE Trans. on Information Theory*, 49:1181–1203, May 2003.

[67] S. S. Pradhan, R. Puri, and K. Ramchandran. n -channel symmetric multiple descriptions Part 1: (n; k) sourcechannel erasure codes. *IEEE Trans. Inf. Theory*, 50(1):47–61, January 2004.

[68] S. Sandeep Pradhan, Suhan Choi, and Kannan Ramchandran. A graph-based framework for transmission of correlated sources over multiple-access channels. *IEEE Trans. Inf. Theory*, 53(12):4583–4604, 2007.

[69] R. Puri, A. Majumdar, and K. Ramchandran. Prism: A video coding paradigm with motion estimation at the decoder. *IEEE Transactions on Image Processing*, 16(10):2436–2448, 2007.

[70] R. Puri, S. S. Pradhan, and K. Ramchandran. n-channel symmetric multiple descriptions Part 2: An achievable ratedistortion region. *IEEE Trans. Inf. Theory*, 51(4):1377–1392, April 2005.

[71] J. Rissanen and M. Wax. Measures of mutual and causal dependence between two time series. *IEEE Transactions on Information Theory*, IT-33(4):598–601, July 1987.

[72] J. P. M. Schalkwijk and T. Kailath. A coding scheme for additive noise channels with feedback-i: No bandwidth constraint. *IEEE Transactions on Information Theory*, 12:172–182, April 1966.

[73] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[74] C. E. Shannon. The zero-error capacity of a noisy channel. *IRE Transactions on Information Theory*, IT-2:8–19, 1956.

[75] C. E. Shannon. Two-Way Communication Channels. *Proceedings of the Fourth Berkeley Symposium on Math. Stat. and Prob., Berkeley, CA*, 1:611–644, 1961.

[76] Y. Steinberg and S. Verdú. Simulation of random processes and rate-distortion theory. *IEEE Transactions on Information Theory*, 43(1):63–86, January 1996.

[77] S. Tatikonda. *Control Under Communications Constraints*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2000.

[78] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *Submitted to IEEE Trans. Info Theory*, arXiv.org:cs.IT/0609139, 2006.

[79] I. P. Tsaregradskii. A note on the capacity of a stationary channel with finite memory. *Theory of Prob. and its Appl. (Russian)*, 3:79–91, 1958.

[80] K Vasek. On the error exponent for ergodic markov sources. *Kybernetica*, 16(3):318–329, 1980.

[81] S. Vembu, S. Verdú, and Y. Steinberg. The source-channel separation theorem revisited. *IEEE Trans. Inf. Theory*, 41(1):44–54, January 1995.

[82] R. Venkataramanan and S. S. Pradhan. Directed Information for Communication problems with Side-information and Feedback/Feed-forward. *Proc. Allerton Conference on Communication, Control and Computing*, 2005.

[83] R. Venkataramani, G. Kramer, and V. K. Goyal. Multiple description coding for many channels. *IEEE Trans. Inf. Theory*, 49(9):2106–2114, Sep. 2003.

[84] S. Verdú and T.Han. A General formula for Channel Capacity. *IEEE Transactions on Information Theory*, 40(4):1147–1157, July 1994.

[85] T. Weissman and N. Merhav. On competitive prediction and its relation to rate-distortion theory. *IEEE Transactions on Information theory*, IT-49(12):3185–3194, December 2003.

[86] F. M. J. Willems and E. C. Van der Muelen. Partial feedback for the discrete memoryless multiple access channel. *IEEE Trans. Inf. Theory*, 29(2):287–290, March 1983.

[87] H. Witsenhausen. On source networks with minimal breakdown degradation. *Bell Syst. Tech. J.*, 59(6):1083–1087, July-August 1980.

[88] H. S. Witsenhausen. On the structure of real-time source coders. *Bell syst. Tech. Journal*, 58:1437–1451, 1979.

[89] A. Wyner and J. Ziv. The Rate-Distortion Function for Source Coding with Side Information at the Decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, January 1976.

[90] H. Yamamoto and K. Itoh. Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback. *IEEE Trans Inf Theory*, IT-25(6):729–733, November 1979.

[91] S. Yang, A. Kavcic, and S. Tatikonda. Feedback capacity of finite-state machine channels. *IEEE Transactions on Information Theory*, 51(3):799–810, 2005.

[92] Z. Zhang and T. Berger. New results in binary multiple descriptions. *IEEE Trans. Inf. Theory*, IT-33(4):502–521, July 1987.